PROTECTING PRIVACY AND DATA IN THE **INTERNET OF THINGS**

Considerations and techniques for big data, machine learning and analytics

FEBRUARY 2019

The GSMA represents the interests of mobile operators worldwide, uniting more than 750 operators with over 350 companies in the broader mobile ecosystem, including handset and device makers, software companies, equipment providers and internet companies, as well as organisations in adjacent industry sectors. The GSMA also produces industry-leading events such as Mobile World Congress, Mobile World Congress Shanghai, Mobile World Congress Americas and the Mobile 360 Series of conferences.

For more information, please visit the GSMA corporate website at www.gsma.com

Follow the GSMA on Twitter: @GSMA

About the GSMA Internet of Things Programme

The GSMA's Internet of Things Programme is an industry initiative focused on:

- **COVERAGE** of machine friendly, cost effective networks to deliver global and universal benefits.
- **CAPABILITY** to capture higher value services beyond connectivity, at scale.
- **CYBERSECURITY** to enable a trusted IoT where security is embedded from the beginning, at every stage of the IoT value chain.

By developing key enablers, facilitating industry collaboration and supporting network optimisation, the Internet of Things Programme is enabling consumers and businesses to harness a host of rich new services, connected by intelligent and secure mobile networks.

Visit gsma.com/iot to find out more about the GSMA IoT Programme.

TABLE OF CONTENTS

01 Introduction

The rapid growth of the Internet of Things creates substantial opportunities for users to benefit from services which are based on data acquisition and storage, analytics and machine learning. It is clear that a responsible approach must be taken by service designers to ensure that personal data and privacy is protected for users, particularly as the IoT expands further into everyday life. This document covers a range of tools and techniques which can be applied to IoT applications and services, particularly where big data storage, analytics and machine learning are employed.

Undoubtedly, one of the biggest advantages of the IoT is the ability to correlate a variety of different data sources to create new services. Using different sources of data allows both users and service providers to gain better insights in relation to a specific situation. For example, disaster and response services that employ IoT and big data technologies, such as real-time, historic and forecast information on weather, water levels, crowds, traffic and other relevant data points, can greatly improve the ability of search and rescue teams to better prepare and react to events. This can be seen in current disaster and emergency response where the loss of life has been greatly reduced by improvements in communications and the distribution of information.

Collection and storage of data through IoT devices should of course meet legal requirements. The GSMA document "Assessing regulatory requirements of privacy management for members offering IoT services using personal data"[1] reviews the requirements from regulators in the EU, US, Japan, India and Brazil regarding privacy of personal data, with these likely to shape other regulatory regimes around the world.

Respecting and protecting privacy is an opportunity to build consumer trust. In "Mobile Privacy Principles"[2], the GSMA has already identified a set of eight principles which help promote consumer privacy in the mobile ecosystem. Some of these principles, such as openness and transparency, data minimisation and respecting user choice and control are particularly relevant to IoT.

[1] https://www.gsma.com/iot/iot-knowledgebase/assessing-regulatory-requirements-of-privacy-management-for-members-offering-iot-services-using-personal-data/
[2] https://www.gsma.com/publicpolicy/wp-content/uploads/2016/02/GSMA2016_Guidelines_Mobile_Privacy_Principles.pdf

GSMA research[3] conducted in the past on mobile services highlights how mobile users really want to manage their privacy in a clear, simple and unobtrusive way. Privacy fears can hold back growth of mobile services and the same logic is applicable to IoT. In essence, end-users really value the way privacy is addressed and building trust with them is a win-win: it drives adoption and enables innovation.

On the contrary, failure to address privacy issues lead to consequences like damage to reputation, or costly law suits; hence businesses will benefit from undertaking a privacy impact analysis before developing and deploying a new service. An example of such lack of attentiveness to privacy led to the Dutch government being forced to revoke its smart metering bill in 2009 due to privacy concerns[4].

1.1 PRIVACY-BY-DESIGN

Designers and service providers can better protect both their end users as well as their own organisations if they consider and address privacy risks as part of service development, implementation and operation. The commonly used expression of 'privacy-by-design' refers to the idea that service designers should focus on privacy throughout the service design process. Privacy is not then something that should be considered at the last moment as a service is launched, or after an issue is identified with a live service. Implementing privacy-by-design, however, can be subjective and will depend on the views and experience of the service designers. To this end, the GSMA has published an "IoT privacy-by-design decision tree"[5] which helps frame the process in a logical and robust way.

This document describes a range of tools and techniques that practically support the concept of privacy-by-design for big data, analytics and machine learning based services in the IoT. It is complementary to previous GSMA publications mentioned above and also the "GSMA IoT Security guidelines"[6], which set out best practice approaches to protect the security of IoT devices and therefore protection of personal information acquired by those devices.

Section 3, in particular, provides approaches to various topics which enhance privacy. Included in section 3 are case studies from a number of mobile operators, illustrating the adoption of such best-practice approaches. This covers the following types of application:

[3] https://www.gsma.com/publicpolicy/wp-content/uploads/2014/02/MOBILE_PRIVACY_Consumer_research_insights_and_considerations_for_policymakers-Final.pdf
[4] See: https://pure.uvt.nl/ws/files/1477311/CPDP_final_Cuijper_Koops_springer_1_.pdf
[5] https://www.gsma.com/iot//wp-content/uploads/2016/09/IoT-%E2%80%98Privacy-By-Design%E2%80%99-decision-tree.pdf
[6] See: https://www.gsma.com/iot/iot-security/iot-security-guidelines/

APPLICATION TO PRIVACY-BY-DESIGN	TOPIC/ SOLUTION	SECTION
Securing data in transit between ✚ IoT devices and IoT platforms ✚ IoT platforms and analytics / machine learning platforms ✚ Analytics/ machine learning platforms and users or partner systems Securely storing data on IoT platforms. Securely storing data on analytics/ machine learning platforms.	Encryption see also Hosting	3.1 3.13
Obscuring data including passwords, items of personal data or commercial data.	Hashing	3.2
Ensuring items of personal data are 'de-identified' in a way that aims to be totally irreversible but prevents correlation of records through the de-identified data. For example, to support analytics and machine learning over aggregated data with a high level of protection of privacy.	Anonymisation	3.3
Ensuring items of personal data are 'de-identified' in a way that aims to be totally irreversible but supports limited correlation of records through some degree of persistence of de-identified data. For example, to support analytics and machine learning over current and/or historical 'de-identified' data.	Pseudonymisation	3.4
Generation of analytics which summarise findings over groups of users or devices including over geographical groupings, types of devices, and general attributes of users. Improving privacy for individuals by dealing with them as a part of a group with similar attributes.	Aggregation	3.5
Checking that outputs generated from analytics and machine learning maintain anonymity of individual users through means such as limiting the 'population' size that can be reported on.	k-anonymity	3.6
Checking that analytics cannot be manipulated in such a way that the results of different queries do not inadvertently disclose attributes of individual users.	Differential privacy	3.7

APPLICATION TO PRIVACY-BY-DESIGN	TOPIC/ SOLUTION	SECTION
Informing users about the service they are using, the reason for collecting and processing any personal data. Providing users with controls governing how their personal data will be used and shared.	Transparency, choice and control	3.8
Ensuring personal user data is not retained beyond the period needed to support the service or services contracted by users. Supporting 'right to be forgotten' for users.	Data erasure	3.9
Ensuring there is clear traceability regarding how data is being used within systems, transferred between systems, and used for purposes such as analytics and machine learning. Supporting compliance for data protection including personal data erasure, as well as contractual/ licensing terms for third party data used in analytics and machine learning.	Data ownership traceability see also Data erasure	3.10 3.9
Confirming with the end-user that they have positively consented to key actions such as logging in to a system, consenting to their personal data being used in analytics, or agreed to a financial transaction.	Two factor/ multi-factor authentication	3.11
Supporting mechanisms to facilitate the controlled flow of data, i.e. the storage or processing of personal data outside of the respective user's country or region.	Cross border data flows	3.12
Guarding the infrastructure that supports services (e.g. data collection, analytics and machine learning platforms) to protect privacy.	Hosting	3.13
Protecting personal and commercial data during ongoing systems development, service delivery, maintenance and support specifically where some or all of a service is contracted out to third-party suppliers.	Outsourcing	3.14

02 Privacy considerations for the IoT

The IoT encompasses a hugely diverse range of devices, things, applications and solutions and therefore it is not possible to choose a definitive set of considerations shared by or covering all things in the IoT. The topics in this section have been chosen as commonly recurring themes in the IoT that may impact privacy. Therefore, application and service developers should consider these issues when designing solutions.

Typical data that is subject to privacy considerations are those that can identify an individual: identification, location information, profiling data linked to an individual or linkage of different data source related to an individual. This document provides a more detailed view of privacy consideration for the IoT in this section and there is a description of the various tools and techniques to improve privacy in section 3.

2.1 DATA IN THE IOT

The IoT provides a diverse range of attributes that are important to understand when addressing privacy. These generally reflect the nature of big data:

- **VOLUME –** the number of IoT devices and things that will connect to the Internet will number into the many billions, so the number of sources of data will be huge;
- **VARIETY –** the type and nature of data will vary significantly between device types, manufacturer approaches and even models of devices. For example a ‘low end’ home thermostat from one manufacturer might simply report a temperature reading to a cloud service and receive control commands to operate the heating, whereas a higher end device from a different manufacturer might use GPS location to optimise the operation of heating & cooling systems according to the weather, and have a greater degree of autonomy over the control process which requires local storage and processing;

- **VELOCITY** - a significant proportion of IoT devices will have the capacity to generate large data volumes especially when they provide real-time or near-real time monitoring and control services. When coupled with the huge volume of IoT devices and things there will be extremely large resultant data volumes. In contrast, however, there will also be significant numbers of devices which need to communicate only occasionally over the lifetime of the device. Due to these differences it is likely that data will be processed and stored in a quite distributed and hierarchical manner and therefore any data that happens to be personal data may not be stored in a single place. Also, it will not be practical to manually scrutinise every data item, source or application;

- **VARIABILITY** – manufacturers might store and process data differently even for similar data values. For example, one manufacturer might store the email address of the owner in the device and another in the cloud. Or variable names may hold very different types of data between two manufacturers e.g. a 'location' field from one manufacturer might hold GPS data where another manufacturer uses this just for the country of the customer. A platform that was designed initially based around receiving low risk data from the IoT device of one manufacturer could later receive higher risk data from another manufacturer, if the second manufacturer 'emulates' the existing interface but includes high risk data without the platform developers' knowledge there is an issue that the developer may be totally unaware of;

- **VERACITY** – the truth or accuracy of recorded data can vary between devices and manufacturers and 'precision' can therefore have an impact on privacy. For example, if an IoT device sends a real-time stream of 1 metre accuracy GPS data it has the potential to disclose much about the life patterns of the owner, but if the accuracy is 1 km and reported on an hourly basis there would be lesser concerns;

In addition to the above factors, longevity of the data and the lifecycle of the service collecting the data are important for the IoT. Some IoT devices may produce very little data but the lifetime of the device could be very long, of the order of 10-20 years. For these devices there may be changes in both ownership and users, typical examples being metering. Those changes would need to be considered and taken into account in evaluating the privacy implications. As an example, the clothing brand Benetton[7] in 2003 encountered protests against privacy violation when they decided to add RFID tags to all garments, as they had not considered the full lifetime of the object.

A common argument has been that 'storage is cheap'[8] so it is easier to store everything rather than spend too much time thinking about what data should be stored. The principle of 'data minimisation and retention' challenge developers to think holistically about data acquisition and storage[9] and this is particularly relevant to the IoT, however, with the above factors it is also much harder to address. In practice data minimisation should be based on the service that the users are receiving at the current time, as well as keeping reasonable options open for future service evolutions whilst avoiding the collection of data that is unlikely to be relevant to customer requirements.

[7] See: https://www.rfidjournal.com/articles/view?344

[8] See Computer Weekly in 2001: https://www.computerweekly.com/feature/Storage-now-cheaper-than-ever

[9] Personal information must not be kept for longer than is necessary for those legitimate business purposes or to meet legal obligations and should subsequently be deleted or rendered anonymous. For further information see https://www.gsma.com/publicpolicy/wp-content/uploads/2012/03/gsmaprivacyprinciples2012.pdf

2.2 LIMITED USER INTERFACES OR USER EXPERIENCE

Internet connected devices have historically had screens and input mechanisms such as keyboards, mice and touch screens. Whether a user had a PC, feature phone or smart phone there was the ability to inform the user via a screen and receive some feedback from the user e.g. clicking confirmation of terms & conditions, registering the device, installing a related application or controlling certain aspects of applications and services.

The capabilities of IoT devices will vary considerably and whilst some of the more sophisticated devices will have screens and input mechanisms this will not be the case for all. Many IoT devices have been designed for 'deploy and forget', with the intent that no interaction is required with humans and therefore they lack any user interface. The interaction with such devices may occur by means of a management system that can be either central or by means of application in traditional devices such as smartphones, tablets or PCs. The capabilities will range, for example, from:

LOW CAPABILITY

The most basic environmental sensor e.g. a rainfall sensor which is deployed in a remote location and sends data to a data collection platform; 1

LOW CAPABILITY

A connected waste bin which uses a combination of sensors to indicate volume, weight, gases and fire and sends this data to the operations centre of a waste management company; 2

LOW CAPABILITY

A remotely operated water valve that is used to irrigate land based on centrally determined weather predictions; 3

LOW CAPABILITY

An automatic blood pressure monitor provided directly to a patient by a health professional or hospital; 4

MEDIUM CAPABILITY

A home 'smart meter' that sends electricity consumption data to a cloud platform and can have its data rendered to the user by means of an application installed on their smartphone; 5

MEDIUM CAPABILITY

A home security system which collects data from multiple movement sensors, door & window switches, and provides a simple numerical keypad for PIN entry along with a small user display; 6

MEDIUM CAPABILITY

An automatic blood pressure monitor provided directly to a patient by a health professional or hospital; 7

HIGH CAPABILITY

An office building control system which provides a touchscreen display for use by the facilities management team of the building for applications including heating/ cooling system management, perimeter security monitoring (PIR, CCTV, etc.) and access control; 8

HIGH CAPABILITY

A connected car which supports various functions including data logging, engine performance optimisation, fault reporting to a manufacturer cloud, navigation, displaying traffic alerts – with a rich user interface provided by a 'touchscreen'; 9

It is likely that the volumes of IoT devices will be significantly higher at the lower end of the pricing and capability curve, and this will mean those devices are less likely to have a display or direct mechanism for user input. It also cannot be assumed that the customer for an IoT service has a 'companion' device (such as a paired smartphone device) with an interface to provide choice and control options or service oversight. Also, increasingly, IoT devices and services will be purchased by people who might not even have a suitably capable mobile device of their own to provide choice & control or they may be purchased by individuals among the estimated 750 million adults who are illiterate[10] and unable to directly exercise their preferences.

Therefore, providing openness, transparency and notice may be challenging, when the IoT devices being used lack suitable display/ input mechanisms. Specific challenges include:

- Viewing and acceptance of Terms & Conditions – if there is no display/ user input mechanism, or the display of the IoT device is not practical for this purpose;
- Providing informed consent for data processing or storage or related purposes;
- Providing any notification in relation to a changed situation, e.g. expiration of consent, changed terms of conditions, new functionality or purpose of processing added that requires consent, etc;
- Asking users to upgrade their devices to obtain security updates that improve privacy.

Solution designers should therefore consider how to achieve openness, transparency and informed consent for IoT based services.

2.3 OWNERSHIP AND USAGE LIFECYCLE

Many mobile devices, including smart phones and tablets, have one user who uses the device all of the time. These are frequently personal devices and the user will personalise their services accordingly including registering or configuring their account details, accepting Terms & Conditions and configuring privacy related features such as location sharing.

IoT devices may be quite different, however. Though an IoT device such as a personal fitness monitor would be used by a single 'owner' there are many use cases in the IoT where there are more complicated ownership and usage scenarios:

[10] Source. UNESCO 2017 data http://uis.unesco.org/en/news/international-literacy-day-2017

LOW COMPLEXITY

A tractor featuring IoT based sensing and control is leased to a farmer by the manufacturer who retains ownership of the tractor and provides the data storage and processing services;

LOW COMPLEXITY

A smart electricity meter is installed into a h ome or apartment and provides regularly updated energy readings to the energy provider. The occupier/ owner of the home may change their electricity provider;

LOW COMPLEXITY

A house with a permanently installed smart IoT home control system is occupied by several family members who must be able to operate certain functions e.g. heating/ cooling/ security. Different suppliers have access to certain functions e.g. security and the owner can change their supplier of any service independently. On sale of the house the home control system and suppliers are transferred to the new owner;

LOW COMPLEXITY

A house or apartment with an IoT based security & fire monitoring system is owned by a private landlord but rented out to a series of tenants with fixed rental periods;

MEDIUM COMPLEXITY

A connected car is owned by one person but normally driven by various family member or friends;

MEDIUM COMPLEXITY

The same connected car is handed over to a service desk at the car dealership regularly and often deposited with parking valets;

HIGH COMPLEXITY

A medical doctor provides a connected blood pressure monitoring device to a patient at risk of a heart attack to assess the success of treatment options. The same device over its lifetime may be used by several patients.

The owner/ purchaser of the device therefore may not be the user of that device, a device might over time be used by different people. Therefore, questions arise about whether the user needs to provide consent, and whether there's a way for a new or additional user to provide their consent. These questions are compounded if the manufacturer cannot discern whether ownership has been transferred, or if the device is being used by someone new.

The lifecycle of the device should therefore be considered when designing both the device and any services based on the device. The service design should consider topics such as device ownership changes and the effect on privacy, and, consider the ability to requalify consent or delete data at the points where ownership changes.

2.4 MARKET AND ECOSYSTEM COMPLEXITIES

As described above, there are various scenarios where the traditional "one person in control of device and application registration or configuration" principle does not apply. This makes processes such as obtaining consent for data sharing even more difficult, even allowing for user interface constraints for the IoT. The following are roles that can be involved in the delivery and use of a rich IoT service such as a building access control service:

- The owner of the 'thing' into which an IoT device has been installed – such as the owner of an office building;
- The owner of the IoT device(s) – such as a specialised security systems service provider;
- The provider of the communications service which the IoT device uses to send/ receive data e.g. the Mobile IoT network;
- The manufacturer of the IoT device;
- The provider of the service which collects and processes the IoT device data and sends any information/ control to the IoT device – including:
    - The provider of the data storage/ processing infrastructure e.g. Amazon Web Services that this main service uses;
    - The partner company who developed the main application/ service and remains responsible for its maintenance;
- The person, persons or organisation contracting with the owner of the 'thing' into which the IoT devices has been installed e.g., a business contracting for office space, who may include:
    - The workers resident in the office who are permitted access to general parts of the office;
    - Workers who are permitted access to specialised parts of the office e.g., IT department access to a machine room;
    - Office cleaners who are provided by a subcontractor to the business;
    - Visitors to the office who are permitted access to visitor areas;
    - An office management team who can configure the access details;
- Other partner organisations who might process access information, for example 'space planning' to improve the capacity or efficiency of the office layout.

In such a scenario, the various external contractual arrangements and internal obligations may allow for data sharing, but there may not be a unified project where all the parties initially sign up to establish the various permissions, consents and restrictions over data sharing. Similarly, there may not be a fixed term project agreement after which all data is deleted as some of that data may reasonably be needed for compliance purposes for many years after its collection. There might be individual agreements established for specific purposes - e.g. engagement of the 'space planning' organisation – and that project and agreement could cover topics such as data encryption and also the erasure of personal data on completion of the project.

Solutions should therefore be designed so that there is support for the different persons or organisations who use the solution to have their privacy protected in line with their role, consents and issues such as contractual compliance.

2.5 PERSONAL DATA IN IOT

Personal data can be quite hard to categorise for the IoT; in general, the more data and more types of data that are collected, the more likely it is that there will be identifiable personal data as part of that data set. Generally, it can be considered that personal data is any information that clearly identifies or is reasonably identifiable for a particular individual. However, the definition of personal data varies with the context in which it is collected. Data that might not be considered personal by nature might become personal when combined with other data sources that allows it to uniquely identify an individual. Some examples:

- The natural name of a person with no other identifying information may not be personal data unless there is a particularly unusual name. 'John Smith' is sufficiently common that the name by itself has little to be concerned about in some countries, but 'Barack Obama' is somewhat more personally identifiable;
- An email address will often identify an individual by their name and sometimes by the organisation they work for. An email address is unique, generally associated with a single individual and therefore generally can be considered to be personally identifiable data;
- Mobile phone numbers are commonly personal identifiers because they are usually used by a single person. In addition, if other related identifiers are held such as device IMEI or IMSI these are also usually associated with a single person. See the later section on Mobile Connect (see https://mobileconnect.io) for the identity solutions it offers related to mobile phone numbers;
- A physical address which is relatable to a large number of individuals (such as a large office block) is not by itself personal data, but if other information is added such as the name of a person who works there it is very likely to be personal data. Conversely a physical address which identifies a single house or apartment is at greater risk of being personal information;

- A single location position for a device is not generally an issue if no other user information is known, but if the positioning is highly accurate and resolves to a location that is otherwise known to be associated with one or a very few people, or the device continuously reports the GPS position over long periods it is more likely that it identifies a single person or sensitive data;
- Post or zip codes that cover large areas of a country or many thousands of people are not directly personal data, but in sparse areas or with highly precise codes such as used in the UK there is an increasing risk these can be associated with small numbers of individuals;
- A ‘hardware address’ of some sort such as the ‘MAC’ address of a Bluetooth Low Energy fitness tracking device is by itself not usable personal data, but if there is widespread collection of this address across a city it may then be possible to identify a person from their lifestyle patterns e.g. home is at, work is at, visits this gym;
- The age of a person in years is relatively low risk for being personal data but the date of birth is much higher risk particularly if this is linked to other information such as their name;
- The temperature reported from a simple sensor is not, when used in isolation, personal data, but if it is known that the sensor is known to be measuring the temperature of a specific person it is more obviously personal data. Alternatively, if a temperature sensor is installed in a precise location, and it is known who resides at the location of the sensor the same temperature data over time might disclose whether the residents are home;
- IP addresses (of connected devices) may be considered to be personal data in some cases, particularly if there is a linking of other data such as name or email address of the registered user with the same IP address;
- Sexual orientation, religion, ethnicity, medical records and genetic data, financial data are among the data categories considered sensitive personal data under some legal frameworks and this will generally be the case in most circumstances.

It is important to recognise therefore that a single item of data generally cannot be considered in isolation and that therefore it is the linking of several data items together either in a composite record, or as part of a sequence of data (time or geographical) that can lead to personally identifiable information - even from single data items that started out as ‘low risk’. For example, whilst a name is generally not personal data, if the age of the person and the city of residence is added, there is a much higher likelihood this identifies a single person (e.g. first name Elizabeth, surname Windsor, home city London, age 92). This ‘joining of data’ has a direct impact on the processes described below where disparate datasets are joined together for analytics purposes.

It is recommended therefore that efforts are made to reduce the data collected and stored based on what is required to support current and planned solutions. Special attention should be paid to data that is of a personal nature, and methods including encryption, anonymisation, pseudonymisation and aggregation should be employed to protect personal data.

2.6 CONSIDERATIONS RELATING TO BIG DATA & AI

The benefits of big data and artificial intelligence/ machine intelligence are generally better achieved when the data set being analysed is both larger (in terms of numbers of records) and broader (in terms of the number of different items available in each record). This at the surface level conflicts with the goal of data minimisation that is reflected in widely accepted privacy principles although when information is not personally identifiable the conflict may be resolved.

In practice, big data analytics and AI are more likely to be used to find common patterns across large data sets through statistical techniques which aggregate across large numbers of users, devices and data. Therefore, to a great extent, these statistical techniques can be considered to be privacy enhancing techniques when applied properly. For example:

- AI could be trained to pre-boil a connected electric kettle at times of the day that historical analysis suggests are common times that IoT device is used. For a simple service like this, it really isn't necessary to know who the kettle belongs to or where it is located but it is necessary to be able to produce the customised programme for any particular kettle;
- Crop growth data, weather data, rainfall data and fertiliser application data can form an extremely useful data set to help farmers correctly irrigate and fertilise their crops. Big data analytics and AI can be used to determine the optimal irrigation and fertiliser strategies for individual fields based on long term historical data. Since the predictive model has been trained using a large number of samples from different farms and fields across a historical data set, therefore it would be the 'rules' that are learned, rather than needing any raw data to be distributed;
- Substantial advances in health condition diagnosis are being achieved from analysis of data obtained across many patients and in many cases from long term studies often established before it was reasonably predictable what health issues might be identifiable or addressable. AI can analyse the data from large numbers of IoT personal fitness monitors, and whilst there's no need to know the personal identity of the users for this purpose, there is a benefit in knowing other data that might be relevant including the age, gender and coarse location. AI can also be trained to identify rare or 'outlier' conditions such as an unusual but treatable 'gait' that could be identified from pedometer data. This type of data could also be excluded as outlier data to preserve privacy. For long term studies the data might be usefully recorded for years or even decades, and for some conditions high frequency data might be useful, e.g. per minute heartbeat data.

Some of the technical issues are:

- Often data from multiple data sets will need to be linked together, and to do this it is necessary to have fields that are unique to a user or IoT device. Identifying the user or IoT device is usually not needed in the resulting joined data set but there needs to be a reliable and reproducible method for linking the data and this might be derived from a device or personal identifier using a method such as hashing, resulting in pseudonymised data;
- The performance of big data analytics and AI are improved through the availability of large and suitably representative data sets numbering thousands or millions of records. For example, an engine management system that was trained using data exclusively collected in countries with hot climates will not perform well for vehicles used in cold climates;
- Identification of the actual users, or too much precise information about users or devices in a data set is often counterproductive in the use of AI/ machine learning as this may lead to the possibility of 'overfitting'[11] – which is when the machine learning algorithms become overly tuned to specific training data and perform poorly when exposed to new data. Whilst the machine learning model is being developed, it might be useful initially to include as many attributes as possible – which are then selectively pruned out during testing and optimisation to avoid overfitting. Excluding personal or device identifiers from the training data should help avoid overfitting;
- However, being able to run a trained machine learning model on data for a particular user or device offers the ability to identify a very specific action that can be taken e.g., telling the user that they are at immediate risk of a heart attack, or closing a floodgate due to rising water levels. Nonetheless, the widely accepted privacy-by design principle of data minimisation tells us there may be limits to storing historical data linked to a user or device. As such, the benefits that big data and AI can generate do not warrant an open-ended permission to store historical personal data and need to be balanced with privacy requirements.
- Asking users to accept or re-accept long and complex Terms & Conditions for new big data and AI use cases or data collection is not really a solution to informed consent and will result in most users simply clicking through without actually reading the Terms & Conditions or otherwise being put off from using the service. Nor is a good solution based on forcing users to process page upon page of check-boxed permissions or regularly asking users to review page after page of new privacy controls. This is even allowing for the issue of IoT device suitability for presenting information and allowing inputs.

When implementing big data storage and AI solutions designers should focus on reducing the amount of personal data stored or processed. Any data should also be stored securely - ideally using storage platforms with hardware encryption. Anonymisation, pseudonymisation and aggregation can also be applied to protect data at all stages of storage and processing.

[11] See https://elitedatascience.com/overfitting-in-machine-learning

2.7 BENEFIT TO USE CASES FROM 'GLOBAL' DATA SETS

A number of countries have introduced restrictions on the flow of data across borders, stemming from national security concerns, data privacy concerns or the desire to protect domestic markets. These restrictions take different forms, such as requiring explicit consent from citizens or prior authorisation from data protection authorities. More prohibitive rules prevent organisations from transferring any personal data or metadata at all.

The effects of such restrictions are many. For instance, requiring organisations to hold an additional copy of data generated from their activities in a country increases the costs of producing physical and digital goods and services in that market. Costs are increased further when the analysis and processing of data must be conducted domestically in addition to storage. In the context of IoT, restrictions on cross-border data flows may led to major impediments to the development and delivery of products and services:

- Individuals who have personal health monitors can also travel on business or holidays and their device would be reporting data to servers in one region, with processing in another, and additional data potentially stored in additional regions. Services would just not work if there was no ability to process data collected across those multiple regions;
- Many products are transported internationally and across continents and an IoT device deployed in the supply chain for product tracking and environmental monitoring can therefore generate data in multiple countries and regions; at certain points in the journey there could be personal data involved e.g., the name and email address of a customs inspector. Such tracking and monitoring again couldn't work if cross border data transfers were not possible;
- Processing of a global data set acquired from a global customer base using IoT personal health monitors might identify conditions that affect particular ethnic groups that might otherwise be lost in the "noise" if data from only one country or region can be analysed;
- Accident and fault data collected and processed by a car manufacturer from across their global sales is more useful than if regionally based data collection and analysis platforms are restricted to collecting data from that single region. For example, the susceptibility of car brakes having poor performance in extremely cold conditions might only be understood and acted upon by collating and processing data from Canada, Northern Europe and Russia.

Techniques such as anonymisation and pseudonymisation can be applied for data storage and processing so that it is possible to deliver solutions using global data sets whilst minimising the risk to user's personal data. Data should also be transferred between regions using strong encryption and the data stored on platforms ideally using hardware level encryption.

03 Technical implementation

The following are a set of techniques and best practice recommendations that provide various protections around the acquisition, storage and use of personal information. These are relevant technical solutions that can be used to implement 'Privacy by Design'. Many of the techniques can be used in combination to support best practice for big data and machine learning implementations.

3.1 ENCRYPTION

Encryption of data, so that it can only be decrypted by a person or a system with a corresponding decryption key, is an important technique to protect data, particularly if it contains personal or commercially confidential data. Unlike 'hashing' (below) there is a defined process to recover the original data using a decryption key.

Encrypted data, however, cannot be processed in analytics or machine learning without it first being decrypted. Both high strength decryption keys and strong security of the decryption key is important to ensure in system design and operation.

Key areas of encryption best practices include:

- Use of a **'Secure element'**[12] in devices for storing certain sensitive e.g. user personal data or user keysets. This is often a dedicated hardware element in the device which is resistant to hacking; for devices that are using the mobile network the SIM is the most appropriate place to store such information.
- **Endpoint encryption** i.e. the encryption of information on a device or removable media so that any stored data cannot be read by a third party unless they have the decryption key;

[12] See https://www.justaskgemalto.com/en/what-is-a-secure-element/

- **Transport encryption** e.g. using the ‘secure’ form of the ‘HTTP’ protocol, ‘HTTPS’, to prevent against man in the middle attacks, for example when transferring data between a device and cloud systems;
- **Encryption at rest** – generally on any system that is storing any data from an IoT device or about the user, or for the storage of data sets before or after analysis or machine learning. This also helps to protect data from recovery by a third party in the case of resale of redundant computer equipment, although this does not mean it will necessarily be secure in the future (see page x);
- Use of strong **encryption measures** e.g. AES-256 to protect data and communications against attack;
- Maintaining **confidentiality of encryption keys**, particularly private keys, especially if partner organisations will need access to those keys for any reason.

There are extensive security recommendations published by GSMA as part of the IoT Security Guidelines at https://www.gsma.com/iot/future-iot-networks/iot-security-guidelines/, these cover topics in encryption both for endpoints and service design.

The ‘Open Web Application Security Project’ provides a very useful set of guidance on best practices for encryption at https://www.owasp.org/index.php/Guide_to_Cryptography

CASE STUDY : China Mobile SAFE link

The IoT introduces an expanded range of security risks over a wide range of technologies including the physical layer (sensors and terminals), transport layer (mobile or other communications networks) and the application layer (traditional Internet security). As an increasing range of industries including manufacturing, agriculture, medical treatment and transportation and logistics use the IoT, the impact of the security risks grows ever greater.

To address these risks China Mobile have introduced a dedicated solution ‘SAFE link’ which combines a ‘security capability cloud platform’, access device security functionality and a security management application. This incorporates traffic filtering, data encryption transmission (supporting quantum encryption), distributed sharing and remote storage and other security capabilities. Security is addressed through capabilities such as secure network access, security monitoring and secure transmission. IoT devices effectively become trusted nodes, the trusted network boundary is pushed to the near data source side and therefore close to users and devices.

SOLUTIONS

The China Mobile solution enables the device to become a secure network access point connecting to the Internet or an Intranet. The combination of device level and cloud platform capabilities enhances the security of the user’s network access request, including interception of malicious URLs, phishing websites, etc, and enables user alerts. The functionality comprises

- **Security capability cloud platform:** This platform integrates a national authoritative security database, ‘Certificate Authority’ system, a China Mobile security database, third-party security database and other security resources. The platform also maintains unified security technical standards and product specifications. Big data processing and analytics technologies achieve security services without user awareness. The cloud platform is designed to address security risks by providing a trusted computing platform, which itself establishes a systematic platform for security protection measures. There are three levels of system management, security management, and audit management supported by the platform. The platform also features proactive defence measures to guarantee the security of the platform itself.

- **Security access equipment:** This specialised equipment co-operates with the device and security capability cloud platform to maintain secure operation of the end user device. Working with the security resources managed by the security capability cloud platform, it monitors URLs requested by the device and handles appropriately requests for non-compliant URLs. The security access equipment also supports commercial VPN/ quantum encryption VPN which enable a secure data transmission channel, supports file encryption and storage sharing, provides users and applications with a high security and high convenience data storage and transmission environment, and extends the mobile security trusted boundary to secure access devices.

- **Security management application:** The security management application can perform various functions such as device management, network configuration, security capability display and remote data storage management on the access device according to the capabilities of the secure access device. The communication with the cloud platform and the secure access device is based on transport encryption which ensures the security of data transmission between the terminal, the security capability cloud platform and the access device.

CONCLUSIONS

The solution establishes protection for border nodes through collaboration between boundary nodes and the security capability cloud platform. This expands the range of defence mechanisms possible for IoT and other applications.

CASE STUDY : Turkcell Cryptographic Key Management

TURKCELL

Enterprises have long acknowledged the necessity of managing personally identifiable information in a secure and trusted environment. Cryptography methods, especially encryption, is increasingly important for protecting such data from disclosure to unauthorized parties. To be able to effectively utilize protection methods in a large enterprise, the most critical operation is the management and use of cryptographic keys, which will protect data for all of its lifetime as well as allow recovery 'by only the authorized personnel' of the related data.

KEY MANAGEMENT SYSTEM

All the keys protecting personally identifiable data must protected against unauthorized disclosure, misuse or modification. Today encryption techniques are well standardised, and their level of security is accepted, however, attention to key management is often overlooked even in large enterprises.

For this purpose, Turkcell have implemented a comprehensive key management system (KMS). A secure key management lifecycle database has been designed and this enables keys to be distributed across the enterprise securely and with full monitoring. Key generation, distribution, replacement and archive operations are supported and are managed through this system.

Every key registered to the key management system is managed as a distinct item with the storage of all related information such as generation/expire time, key check values, and usage. Also registered keys can have broader relationships with other system's configuration items (e.g. database, server) which makes it easy to keep track of where keys are used and their operational state.

When there is a requirement to set up a new application or system that uses encryption or digital signatures, the key generation process is started via the submission of a new ticket to the key management system. The owner of the application service submits all the necessary information such as key type, algorithm, length and key storage type and a new key is generated with the approval of security teams. The entry is created in the KMS database along with all relations to the various distributed systems which will use that key.

If there's a need to provide a key to a third party via an external interface, custodians are selected and assigned with tasks on this system. This aids security teams to easily log key distribution online.

Lastly, key replacement operation is also automated, starting with notifications to the service owner prior to the expiration date. The service owner decides if a new key should be generated and updates the notification ticket, which triggers key generation process. Previous keys are archived on expiry or update and related configuration information is flagged as non-operational.

CRYPTOGRAPHIC WEB SERVICE

Cryptographic keys need to be available locally at the various systems and applications (e.g. web servers) which users or systems rely on for encrypting, decrypting or signing in to a service. This is necessary throughout the key's active lifetime. Typically, operational storage is selected for the environment which uses the key (e.g. server disk storage, or hardware security module). Of special importance is the implementation of security controls for data encryption and private keys. It is also important to select appropriate cryptography algorithms at the application level to suit security needs. These two factors are often seen as a burden by teams outside of the security operation and are challenging for web service designers both when developing new applications and in maintaining the application as keys are updated, new cryptographic algorithms are introduced and older algorithms retired.

Turkcell have designed an improved architecture for applications where encryption, decryption and key management is provided as an internal "cryptographic web service" used by applications. Each application is individually identified and authorized to use keys generated for that application. The application never actually accesses any key material, all of the operations which require the use of keys are provided externally to the application via the cryptographic web service. Furthermore, a hardware security module is used by the cryptographic web service to store all the key material, and cryptographic methods are run on dedicated server applications.

This architecture allows isolation of key material and related cryptographic functionality from the application since the application only needs to login, provide data, and read the output from the cryptographic web service. Key storage, replacement, retirement and archival is handled by the cryptographic web service and the application exposure to key lifecycle management is minimised e.g. notification of a key replacement operation.

3.2 HASHING

There are situations where it is useful to be able to hold information "secret" in a coded way without ever needing to decode the coded information. A common example is in the way passwords are often stored on systems; as an example, if a password comprises the digits 65349811 the hash function result[13] of this is 4A12751EE967410334811ECE84FB16FA and so any password entered which doesn't have the same hash value is known to be incorrect without adding to the increased risks involved in storing the actual password.

Hash values are useful because they are highly resistant to reverse engineering, especially when strong hashing algorithms such as SHA-256[14] are used. For example, if a hash value is generated for an email address an attacker would have to attempt the hashing process for all known or likely email addresses in order to see if the user existed in a particular data set and this is generally not worth the time or effort. Hash values are therefore useful to obscure personal information in data sets, but in a way that is deterministic i.e. the hash value of a given input will be the same for a given hash algorithm. Hash values can therefore be used to confirm something that another system or a user knows and in big data/ machine learning provide a useful tool for hiding personal data in a way that protects privacy. The hashing function is also a useful technique to apply for pseudonymisation.

Best practice recommendations for hashing are that secure hashing algorithms (e.g. SHA-256 or higher) are used, along with salting to prevent dictionary attacks against hashed data.

3.2.1 SALTING

This is a process where random data is added to other data before processes such as cryptographic hashing are applied. This is a technique used widely in the storage of password values and helps to avoid correlation of passwords which have been cracked on one system being used to reverse engineer passwords stored on another system.

The particular problem being addressed is that hashing functions, by necessity and design, produce the same output bytes for the same input bytes. Therefore, if a system uses a specific hashing function without adding other data (the 'salt') to the input value it is possible to compare hash output values for equality. There is a comparable vulnerability for encryption processes that is addressed using an 'Initialization Vector'[15].

[13] This is using the MD5 hash function, though this is no longer considered robust and more secure hash algorithms such as SHA-256 should be used
[14] SHA-256 was designed by the US NSA. For more information see https://brilliant.org/wiki/secure-hashing-algorithms/
[15] See https://csrc.nist.gov/Glossary/?term=5027

‘Salting’ has broader application in privacy because, for example, if a specific hashing function is applied to an email address that will have a distinct hashing signature that would be the same on any other system or in any other dataset that uses the same hashing function on the same email address. It would be possible to correlate such data by searching for the same hash values. Additionally, for some data with relatively limited possibilities it is feasible to perform an attack such as a dictionary attack[16] using a set of the most likely values for a hashed field. For example, determining a date of birth from a hashed value requires only the processing of around 43,800 hashed values (120 years x 365 days of the year[17]) – this is a method known as a ‘brute force attack’[18] and if the computed hashed values are pre-generated this is a dictionary attack.

Salting requires the random value to be selected and applied to the input data being hashed. For example, in the date of birth case a randomly selected integer could be added to the year part of the date, with a different random offset (the ‘salt’) generated for each record in the input data. Using a different salt value for each record further protects against attempts to crack the hashing using brute force and dictionary attack methods. In the ‘date of birth case’ adding a salt value in the range -500,000 to +500,000 to the year part would mean the number of hash values to attempt would be 365,043,800.

This process ensures that for any data item subject to hashing, there is reproducibility in the generation of the hash for given input data for purposes such as historical trend analysis, but there is a protection of personal data against viable attack methods. Ideally the salt values should be stored separately from the actual data records and stored on a volume using strong encryption.

[16] See https://searchsecurity.techtarget.com/definition/dictionary-attack
[17] Ignoring for this example leap years
[18] See https://searchsecurity.techtarget.com/definition/brute-force-cracking

3.3 ANONYMISATION

This is the process of converting data, often personal data or data that might conceivably become personal data when subsequently processed, to a form that is irreversible and impossible to relate to the original personal data. The Future of Privacy Forum[19] defines such 'de-identified data' in the following ways:

i. data from which direct and indirect identifiers have been permanently removed; or

ii. data that has been perturbed to the degree that risk of re-identification is small, given the context of the data set; or

iii. data that an expert has confirmed poses a very small risk that information can be used by an anticipated recipient to identify an individual.

The purpose of anonymisation is to enable data to be stored or communicated to other systems or organisations for storage or processing, whilst protecting user privacy or other confidentiality (e.g. commercial confidentiality).

There are various methods for anonymisation:

- Simply removing a particular field from the source data, e.g. if it is known that device location is always going to be irrelevant to analytics, then it may simply be removed -either by removing the whole field or by replacing by a null value across all records;
- Removing whole records which contain personal data, e.g. any record where the location is near a religious place of worship[20];
- Replacing a particular field with a standard substitute value that means it has been anonymised. E.g. '*' for a text field, -999 for a numeric field (provided that field is never expected to have such a value);
- Replacing a specific field with a totally random replacement, e.g. a random text string value or a random numeric string value;
- 'Character masking' applied to textual values can be applied so that the sensitive data is sufficiently obscured to no longer be personal. For example, the name 'Alice Smith' could be changed to 'A* S*'. This is commonly used with credit card numbers where the number is stored by obscuring a number of the digits e.g. '1234 **** **** 8901' or '**** **** **** 8901'. This technique can be useful if there is a requirement for displaying the anonymised field to the user at a later point but there is also a risk that a masked field could be used for correlating;

[19] See https://fpf.org
[20] Note that this may not be a particularly robust method of anonymisation as an absence of certain records could be interpreted in a way that identifies something about a user or device.

Anonymisation can be performed when records are received from other systems, or from devices, so that there is a reduced risk on storage of data. Also, anonymisation can be performed after data has been processed so that any storage or communication of results maintains user privacy. The best practice recommendation is to anonymise data at the earliest opportunity possible to prevent unintended access to the original data.

One important disadvantage of anonymisation is that it will generally restrict the types of data processing and analytics that can be performed on the data. For this reason, other methods including pseudonymisation are often preferable to ensure better use can be made out of the data whilst still maintaining user privacy. However, many data protection laws, such as the EU GDPR, do not apply to anonymised data. Even so, techniques such as differential privacy should be used to avoid privacy risks when anonymised data sets are joined together.

3.4 PSEUDONYMISATION

To protect individual privacy a key solution that can be used to turn personal data into 'de-identified' data is pseudonymisation. In this, an original piece of personal data is transformed, usually with a one-way mapping table or function, into a value that is unique for the original data but irreversible to a third party.

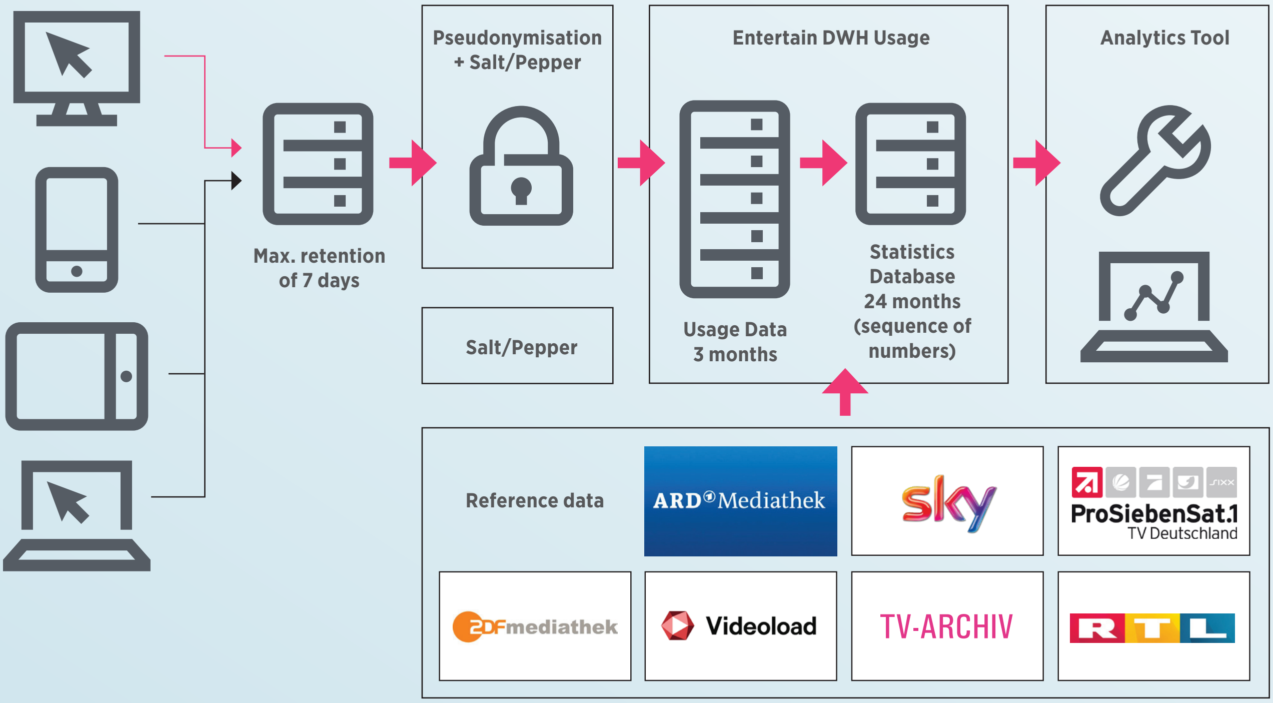
Pseudonymisation is generally achieved by the substitution of an alternative, system generated identifier to replace one or more fields of personal data of the user. If the pseudonymised identifier is part of a resulting data set, or shared with external parties, it should be impossible for recipients to identify personal information about the user from that data set whilst other analytics remain possible e.g. trend analysis for an IoT system. As with the best practice for anonymisation it is recommended that 'personal data' is pseudonymised at the earliest opportunity during data storage and processing.

CASE STUDY : Pseudonymisation and Entertain TV (Deutsche Telekom AG.)

Deutsche Telekom AG. markets access to television programs and films via the internet under the product name of Entertain TV. The company provides customers with a set-top box for using the product. Statistics on viewers' habits are maintained for various purposes, including obligations vis-a-vis broadcasters.

OVERVIEW

The following diagram shows the data flow through the individual systems, from the set-top box to the statistics. update and related configuration information is flagged as non-operational.



DATA GENERATION

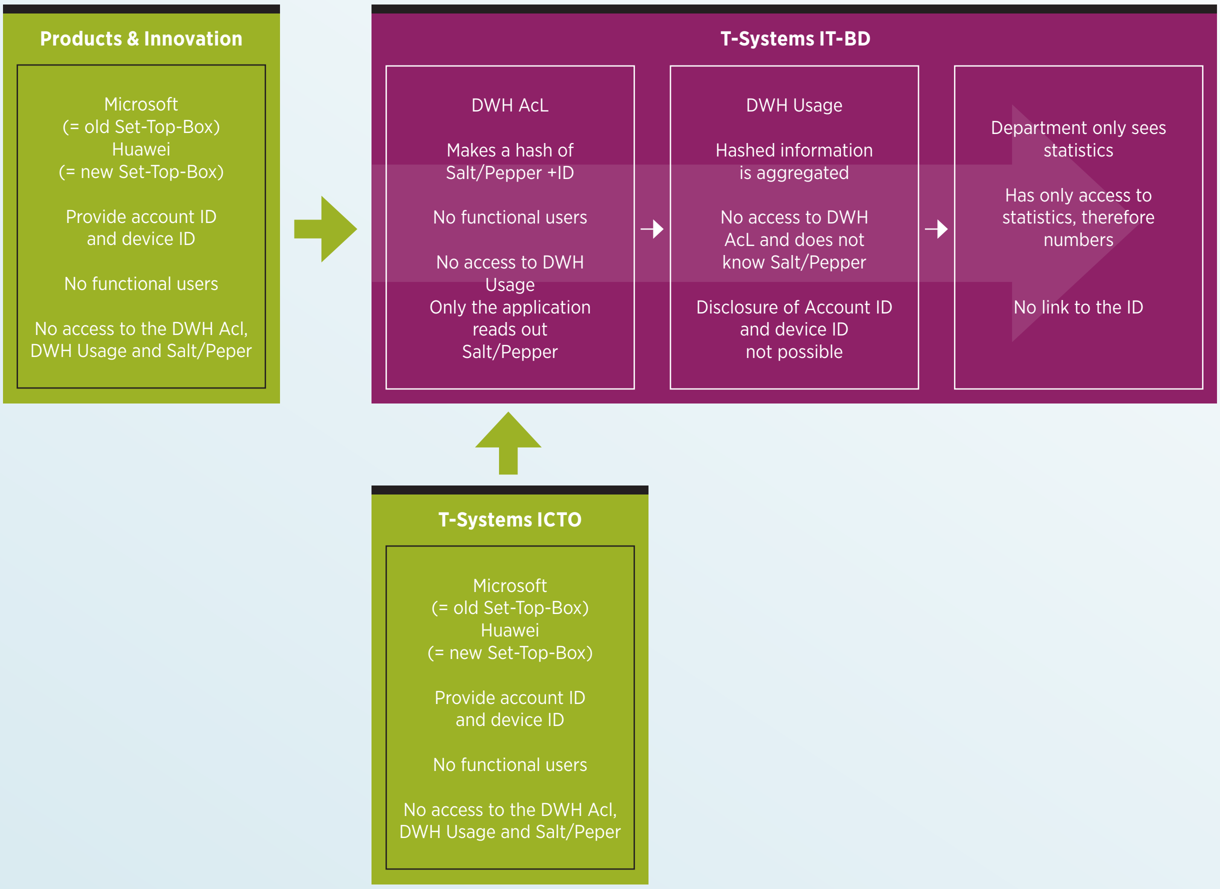
Use of the set-top box, e.g. when the consumer uses the system’s remote control, generates a range of events depending on what button was pressed and the relevant context. These events form the basis of various analyses, which document activities such as activation/deactivation, channel changes, information about the programs watched, information about users’ recording activities, or information about users watching recorded programs. Corresponding event data sets contain a range of information, i.e. about the set-top box (device ID), the customer’s account ID, date/time, and other specific subjects.

PSEUDONYMISATION

The account ID is a pseudonym for the customer, and the device ID is a pseudonym for the associated set-top box. The event data sets required for analysis purposes do not contain any attributes featuring personal data (of an immediate kind). Managed separately in organizational terms, there are allocation tables which permit the pseudonyms (account and device IDs) to be linked to customers or set-top boxes. Access to these tables would make it possible to ultimately trace the device and account IDs to trace back to the customers. Tracing is sometimes necessary, i.e. for billing services as per the contract.

However, as no party wants to trace details back to “plain data” for the purpose of generating statistics, the account and device IDs are subjected to additional pseudonymisation before processing. In this instance, pseudonymisation takes place within the Data Warehouse Acquisition Layer (DWH ACL) unit and processing (statistics generation) takes place in Date Warehouse Usage (DWH Usage), a separate unit (as shown below).

ORGANISATIONAL AND AUTHORISATION CONCEPT



Products & Innovation

Microsoft
(= old Set-Top-Box)
Huawei
(= new Set-Top-Box)

Provide account ID
and device ID

No functional users

No access to the DWH Acl,
DWH Usage and Salt/Peper

T-Systems IT-BD

DWH AcL

Makes a hash of
Salt/Pepper +ID

No functional users

No access to DWH
Usage
Only the application
reads out
Salt/Pepper

DWH Usage

Hashed information
is aggregated

No access to DWH
AcL and does not
know Salt/Pepper

Disclosure of Account ID
and device ID
not possible

Department only sees
statistics

Has only access to
statistics, therefore
numbers

No link to the ID

T-Systems ICTO

Microsoft
(= old Set-Top-Box)
Huawei
(= new Set-Top-Box)

Provide account ID
and device ID

No functional users

No access to the DWH Acl,
DWH Usage and Salt/Peper

The underlying pseudonymisation process means that statistics are generated using pseudonyms that are linkable but non-disclosable for DWH Usage, the unit involved in processing. The pseudonyms are created using a deterministic cryptographic hashing process (SHA-512) and a uniform pseudonymisation salt/pepper component. Pseudonym link-ability is established because deterministic processes transpose identical plaintexts onto identical result values (pseudonyms) when the salt/pepper is identical. The output length of SHA-512 means that the risk of collisions is negligible ($< 10^{70}$). As DWH Usage does not have access to the pseudonymisation salt/pepper, it cannot practically trace pseudonyms back to actual people and so disclose plaintexts.

Ultimately, the pseudonymisation process used means that no Deutsche Telekom Group employee can view, analyse, or transfer to anyone else information about specific customer's usage behaviour.

STATISTICS GENERATION

All attributes traceable to the end customer are pseudonymised using an account or device ID. Payments are possible as linkable pseudonyms are used. For example, this means that it is possible to answer a question about how many households or set-top boxes watched a certain channel at a certain time. The anonymous statistics no longer contain account and device IDs or the generated pseudonyms, which prevents the statistical figures from being traced back to the hashed IDs.

Deutsche Telekom must meet certain obligations towards broadcasters, so it transfers only anonymised statistics on viewers' user behaviour, e.g. market share using relative figures.

OPT-OUT

Data protection notices inform every Entertain TV customer that data is gathered for statistical purposes. Deutsche Telekom uses e-mails and pop-ups in Entertain TV itself to inform customers of this before introducing this analysis solution.

Every customer has the option of objecting (opt out) to their pseudonymised usage data being collected and analysed. The customer can use their set-top box to perform this opt-out. Previously this had involved the input of a PIN number but with newer set-top boxes no PIN entry is needed. By opting out, the customer's usage data is not used either for a pseudonymised usage profile or for anonymous statistics. Customers can also use conventional communication channels to inform Deutsche Telekom that they want to exercise their opt-out right.

Pseudonymisation has certain advantages over methods such as anonymisation and aggregation:

- It is possible to maintain a per-user or per-device data set, just without the data set disclosing the identity of the user or device;
- It is possible to associate multiple data sets together if there is a common pseudonymous identifier appearing in those data sets, without needing original personal identifiers;
- If the analytics or AI identifies an issue for that user or device, it is still possible to implement an action or response that can be provided to that user or device – if there is the support of a higher-level entity that knows the mapping between original identifiers and pseudonymous identifiers;
- It is possible to maintain a continuous thread covering both historical and current data so that analytics can be conducted across the time domain e.g. trend analysis;
- More personalised IoT services can be delivered, even though the identity of the user remains better protected.

Best practice recommendations for pseudonymisation include the use of one or more of the following:

- Use of mechanisms such as a long, random UUID's (Universal Unique Identifier) to substitute for one or more pieces of personal data. At least a 128-bit Type 4 random UUIDs are recommended e.g. 'cb3eb75f-cbbc-414c-a146-624518ed7537' (hex coded);
- Use of strong cryptographic hashing[21] functions (e.g. SHA-256 or better) with 'salting' to one-way encode personal data in a way that cannot practically be reversed;
- Generating different pseudonymous identifiers for personal data for each separate application, including analytics or machine learning process, that uses the data;
- Generating different time bound pseudonymous identifiers for personal data over defined time periods e.g. a new pseudonymous identifier is generated each day for any given personal data field and any given application so that any downstream application cannot track the long-term actions of users[22].

In the case that a service may need to use the result of analytics to generate a result or alert for a user or device that is identified using a pseudonymised identifier a more secure implementation can be produced by separating the platform[23] which stores the personal data from the platform that conducts the analytics:

[21] See https://www.sans.edu/cyber-research/security-laboratory/article/hash-functions
[22] This technique protects against building up a long term historical data set for a user/ device which could risk becoming personal data by longer term analysis of usage
[23] Physical and/or logical separation and with different access controls.

- The platform holding the user or device data (the 'master data' system) maintains the knowledge of how to 'reverse map' from a pseudonymised identifier back to the original user or device;
- The analytics platform passes the pseudonymised identifier to the master data system along with any result or action, and the master data system then handles the routing to the user;
- This mapping data store must be highly secure, have the most limited access possible to both administrative staff and connected systems, and employ strong encryption for any stored data;
- If there is no need to offer such services, there is no need to deploy such a reverse mapping store;
- If an application or service is retired the mapping store should also be purged of the data relating to that application or service;
- If a user terminates their service, the mapping store should also be purged of the data relating to that user.

3.5 AGGREGATION

Personal data is better protected if data from multiple users or devices is aggregated in some way. This is because the data on a specific individual is then 'hidden' within a more general cohort. An example of various practices employed by the US Census Bureau is published in the paper "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research"[24] and there is further useful guidance in the U.S. Federal Committee on Statistical Methodology publication "Report on Statistical Disclosure Limitation Methodology"[25].

There are various aggregation techniques that help improve personal data privacy. Some of the techniques can be employed on data before it is subject to additional processing and analytics and some of the techniques are useful to apply, post analytics, to ensure that users do not become recognisable in the results.

[24] See https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf
[25] See https://nces.ed.gov/FCSM/pdf/spwp22.pdf

The techniques that are covered here are:

- Generalisation of personal information;
- Time domain aggregation;
- Geographical domain aggregation;
- Group level aggregation.

See also the sections on k-anonymity and differential privacy as these provide methods to assess and enhance the privacy of data and results even when aggregation is used.

Note that in all cases the methods to use are dependent on the actual use case. 'Privacy by design' encourages developers to think about how data is used at all stages to best maintain user privacy.

3.5.1 GENERALISATION OF INFORMATION

For some analytics it is necessary to know some information about a user without needing to know very specific personal information. For example, a 'black box' insurance service[26] is normally charged based on the age of the driver, but it is not strictly necessary to know their actual date of birth when determining the insurance premium.

In many cases it is possible to convert discrete values into distinct ranges. This can also assist certain machine learning techniques where continuous ranged variables (such as temperature, voltage, flow rate) must be converted into 'buckets'[27] to apply techniques which combine multiple values together to form a new column ('crossed column'[28]).

One of the simplest examples for this would be converting the date of birth into an age range, which might be useful in personal health analytics. For this the age of the user could be allocated into a range of 0-18, 18-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75-85, and 85+. The technique can also be used for other information such as: the number of vehicles passing through a road junction, height of a person, depth of water, amount of energy consumed in an hour, measured voltage level, etc.

For example, in the following US public data set of census and salary, the columns "Age", "Education Years" and "Hours Per Week" could each be converted to 'bucket' values. Not only does this improve the privacy of the data set, it would also allow a wider range of machine learning algorithms to be applied to the problem of estimating salary[29].

[26] i.e. an insurance service that uses an in vehicle telematics system to monitor driving behaviour
[27] See for example https://www.tensorflow.org/api_docs/python/tf/feature_column/bucketized_column
[28] See https://www.tensorflow.org/api_docs/python/tf/feature_column/crossed_column
[29] See the tensorflow example https://github.com/tensorflow/models/blob/master/official/wide_deep/census_dataset.py

AGE	WORK CLASS	EDUCATION	EDUCATION YEARS	MARITAL STATUS	OCCUPATION	RELATIONSHIP	RACE	GENDER	HOURS PER WEEK	NATIVE COUNTRY	SALARY LEVEL
39	State-gov	Bachelors	13	Never married	Adm Clerical	Not in family	White	Male	40	USA	<=50K
50	Self-emp-nc	Bachelors	13	Married Cicil	Exec-manager	Husband	White	Male	13	USA	<=50K
38	Private	HS-grad	9	Divorced	Handlers-cle	Not in family	White	Male	40	USA	<=50K
53	Private	11th	7	Married Cicil	Handlers-cle	Husband	Black	Male	40	USA	<=50K
28	Private	Bachelors	13	Married-Cicil	Prof-special	Wife	Black	Female	40	Cuba	<=50K
37	Private	Masters	14	Marrie-Cicil	Exec-manager	Wife	White	Female	40	USA	<=50K
49	Private	9th	5	Married-Spouse	Other-service	Not in family	Black	Female	16	Jamiaca	<=50K
52	Self-emp-nc	HS-grad	9	Married-Cicil	Exec-manager	Husband	White	Male	45	USA	>50K
31	Private	Masters	14	Never-married	Prof-special	Not in family	White	Female	50	USA	>50K
42	Private	Bachelors	13	Married-Cicil	Exec-manager	Husband	White	Male	40	USA	>50K
37	Private	Some-college	10	Married-Cicil	Exec-manager	Husband	Black	Male	80	USA	>50K
30	State-gov	Bachelors	13	Married-Cicil	Prof-special	Husband	Asian-Pac-Isl	Male	40	India	>50K
23	Private	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	30	USA	<=50K
32	Private	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	50	USA	<=50K

3.5.2 TIME DOMAIN AGGREGATION

Privacy can be enhanced by adjusting the time domain over which data is aggregated. For example, if a home smart meter reports energy usage every five minutes it may be possible to make an informed guess regarding whether anyone seems to be at home, this is easier to assess if there is also access to a period of historical data also at five-minute intervals. That insight might not be particularly useful unless other personal data is known (such as the actual location of the smart meter), but if the five-minute data is aggregated to hourly, daily or weekly data there is much better protection against a third party determining something more personal about a user such as whether they appear to go out to work.

For some use cases it will be important to use 'real-time' data, this will be so particularly for control applications. For other use cases there may be an advantage in using statistical techniques to aggregate the data e.g. determining a mean value, standard deviation, or minimum/maximum range over a longer period of time.

If a third-party application is given access to analytics with too much ability to influence the period over which data is aggregated it may be possible for it to make varied requests to drill down into the data. This is especially true if the application can request finer grained intervals than strictly needed for its use cases.

3.5.3 GEOGRAPHICAL DOMAIN AGGREGATION

Location of a device or user is important to many use cases, particularly in applications such as personal navigation where very precise 'real-time' location is required for step by step navigation. Location can also be useful for advertising, transport provision, traffic control, retail supply chain optimisation and so on. The user or device location becomes particularly sensitive if it can be related to other personal data.

Precise user location, particularly with retained historical location data, can risk the privacy of individuals:

- By identifying where the user or their connected device/ thing spends significant amounts of their time (e.g. work, home);
- By identifying routes that the user or their connected device/ thing routinely travels;
- By identifying places that the user or their connected device/ thing is routinely visiting – this may be a particularly sensitive issue if for example it is determined the user is visiting a physician or hospital.

The precision of location of a device or user does not have to be the same across all use cases, and there are various best practices that can be employed to mitigate against misuse of location data:

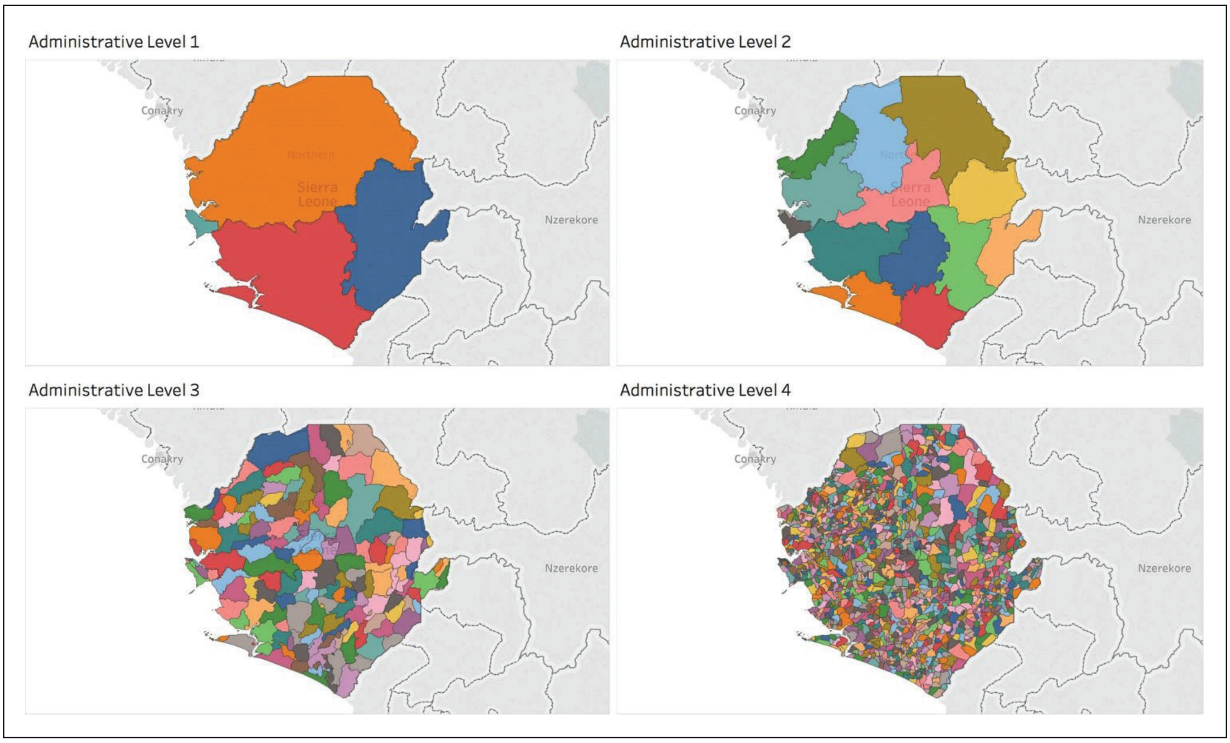
- Decrease the precision of the latitude/ longitude by using a smaller number of significant digits when location data is stored or used in analytics;
- Report the location of a user or device based on a larger geographical grid, for example using the 'geohash' system[30] which supports multiple geographic resolutions;
- Resolve the location of a user or device to a post or zip code area (or partial post code for countries such as the UK which have some very small postcode areas);
- Relate the user or device location to a town, city, district or 'administrative area'. An example of this is shown below for Sierra Leone with four levels of administrative levels;
- Limit the access to historical location data, relevant to the use case.

The 'administrative area' approach can also be useful if combining datasets with government data as that will often be reported for an administrative area. However, there can also be a complex conversion involved that is reliant on government provided spatial data[31].

Note that the location of a user or device is not just sensitive from a personal data perspective, it might also reveal commercially sensitive data about the location of a customer base.

[30] See https://en.wikipedia.org/wiki/Geohash - this has a similar effect to reducing the number of significant digits of the latitude/ longitude

[31] Spatial data is also not always available



Example of different administrative levels in Sierra Leone

3.5.4 GROUP LEVEL AGGREGATION

Personal data is better protected when data from multiple individuals is aggregated into a group. Usually this means that there is either a statistical aggregation (e.g. average height, weight, age) or grouping of people with attribute values in a broader range, or by their location. This form of grouping makes it much more difficult to identify an individual from a set of data or results.

The size of the group influences the degree of privacy for each individual. Therefore, larger groups will generally offer a greater degree of privacy.

Grouping can be applied to input data e.g. by aggregating or analysing the energy usage reported by smart meters across the whole of the customer base there is no longer a sensitivity around the energy usage of a single customer. The customer base could be further segmented into customers in a certain city, or by age range of the customer. As long as the number of members of any selected group remains above a reasonable “floor” level there is a low risk to personal information.

What is important therefore is to establish a sufficiently large group size to protect the privacy of individuals whilst not choosing so large a group that the insight about the group is too 'general'. The best practices used by government agencies producing official statistics are useful. For example, the US census only releases data on race, ancestry and ethnicity provided this is assessed for a minimum of 100 people in a given geographic area[32], it also recommends ages are grouped into brackets of 5 years. The UK Office of National Statistics identifies the following risks in the sensitive area of births and death statistics[33]:

- A cell with a count of 0, 1 or 2 has a potential disclosure risk. The advice is therefore to suppress any output with a count under 3;
- Row and column totals might also present a risk;
- A table with many dimensions, variables or small cell counts is likely to be riskier than one with fewer. In output reports this means it is better to use fewer columns;
- Overlapping geographical areas and rolling multi-year aggregates can easily give rise to disclosure through differencing.

Australia's New South Wales Ministry of Health recommends amongst other guidance in the publication "Privacy issues and the reporting of small numbers"[34]:

- Statistical results involving small numbers can be presented if drawn from a population of at least 1000 people;
- Use either 5-year or 10-year age aggregations depending on use case;
- Do not include counts in cells for users or devices where the total matched is less than 3, 5 or 10 depending on use case;
- Row and column totals (or marginal totals) make it more complicated to protect the privacy in suppressed cells.

[32] See https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyonprotectingconfidentialityintablesofbirthanddeathstatistics
[33] See http://www.health.nsw.gov.au/hsnsw/Publications/privacy-small-numbers.pdf
[34] See https://epic.org/privacy/reidentification/Sweeney_Article.pdf

3.6 K-ANONYMITY

This is a technique for analysing and protecting privacy where there is sufficient uniqueness in the data stored or released that effectively the subject can be identified when combined with external data sets. The technique was described in a paper by Latanya Sweeney[35] with the example that 87% of the population of the US could be uniquely identified from the use of publicly available data sets containing a 5-digit ZIP code, gender and date of birth.

A data set provides k-anonymity protection if each person in a data set cannot be distinguished from at least 'k-1' other individuals in the same data set. 'k' is a factor that is chosen according to the degree of uniqueness required in a data set. k-anonymity is also useful to assess for the data sets used for machine learning as there is a risk that the models will not perform well for new data if they have somehow learned to produce specific results for the unique individuals in a data set ('overfitting').

The specific problem being addressed is that simple anonymisation, pseudonymisation and aggregation techniques focus on removing personal data from individual fields, but do not consider how the information that remains combines together to uniquely associate with an individual ('a digital fingerprint'). If a resulting data set is then combined by a third party with data having the same 'fingerprint' and who also has, for example, the email address, name or home address of the user they can relate the data that was believed to be anonymised to an actual person. K-anonymity can therefore be considered a best practice to review the output of analytics and machine learning to ensure personal data remains protected after data has been processed.

k-anonymity can be assessed for a data set by assessing the number of matches between any record and all other records. The process can be lengthy for large data sets and data sets with many attributes. k-anonymity of the right degree is then achieved by applying, for example, the following methods:

- Increasing the generalisation of a field e.g. taking only the first four digits of a Zip Code, or using a range for a value e.g. age;
- Suppressing fields that create uniqueness to an individual between rows;
- Suppressing records that do not achieve k-anonymity;
- Randomly sorting the result set to avoid inference e.g. that the age of users is in a particular order.

Note that k-anonymisation attempts to address individual privacy through making sure that any one individual is not uniquely identifiable in a data set. However, even so, there remains a risk that a particular user might be one of a group matching particular criteria especially if results are combined with external data sets. Differential Privacy (following) is a more robust mechanism to better protect user privacy.

[35] See https://epic.org/privacy/reidentification/Sweeney_Article.pdf

CASE STUDY : Deutsche Telekom / Motionlogic – application of k-anonymity

Telekom Deutschland GmbH (TDG) is providing anonymised signalling data from the mobile communication network to the company Motionlogic GmbH for onwards analysis for use in transport and other analytics. This is provided for mobile customers of Telekom Deutschland GmbH (TDG). Importantly, and even though their data is only ever used in an anonymised and aggregated form, the TDG mobile customers have the right to opt-out of their data being included within the processing described below.

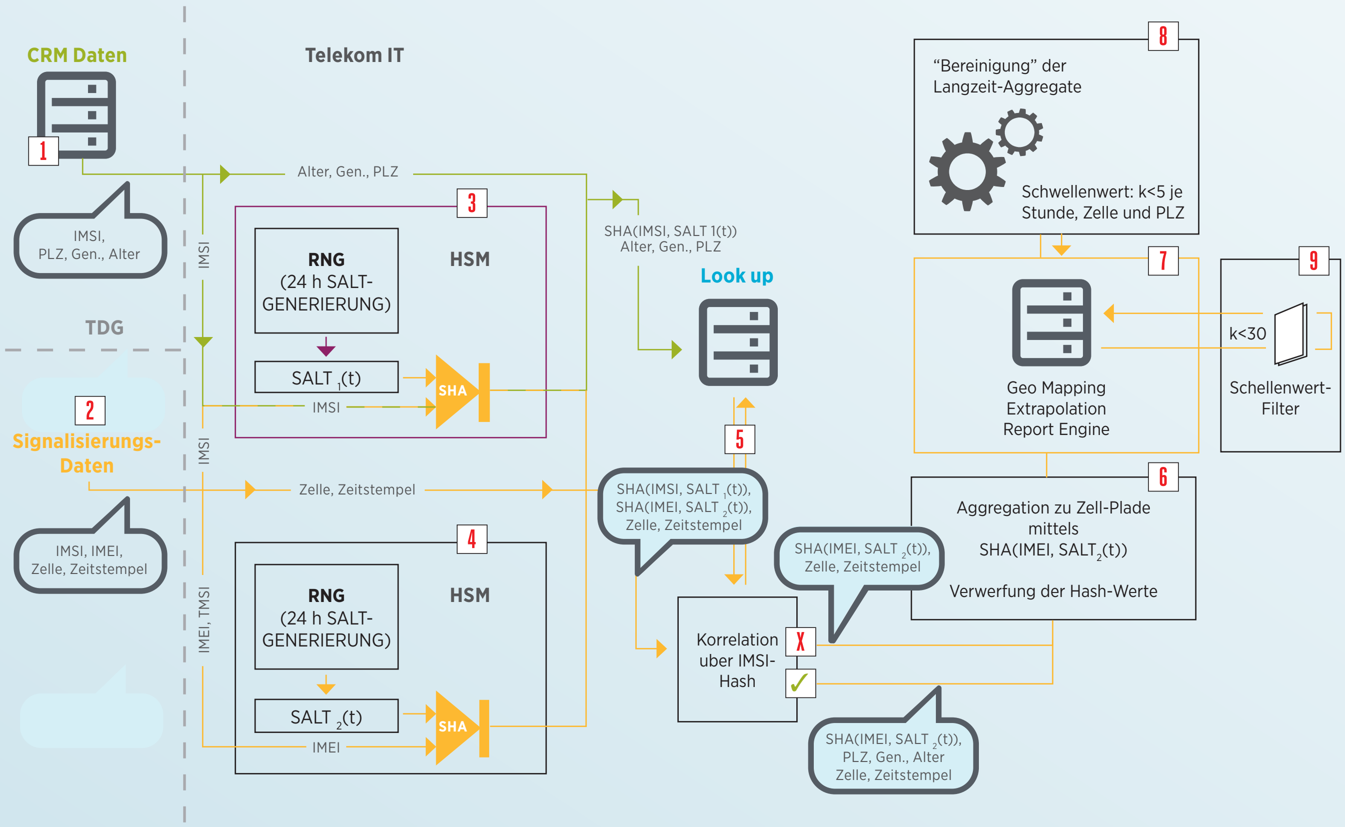
The starting point for the process is mobile network signalling data which indicates cell and timestamp, coupled with high-level user attributes (age in 10-year intervals, gender, zip code) from the Deutsche Telekom Customer Relationship Management (CRM) contract database. As a first protection for this data the IMSI, which allows data to be correlated between the mobile network data and the CRM database, is hashed using a 'salt' value applied identically to the mobile signalling data and the CRM database data to prevent discovery of the IMSI. The 'salt' value used for IMSI hashing is replaced every 24 hours. In addition, the customer's IMEI (equipment identity) is also hashed using a different 'salt' value which also has 24-hour validity period and this is added to the pseudonymised customer data. Deutsche Telekom ensures that data is not and cannot be traced back to individual persons by the following:

- The hash process is carried out using confidential SALTs
  - A SALT change occurs for SALT1 (used for pseudonymisation of the IMSI) and SALT2 (used for pseudonymisation of the IMEI) every 24 hours. After the change, the old values are discarded.
  - SALT values are generated on a pseudorandom basis, and technically created and used in an encapsulated system (Hardware Security Module) without human intervention.

  The respective SALT therefore exists only within this time period and is accessible only for the above-mentioned internal system purposes.
- Hashing the IMEI from the mobile network signalling data using SALT2 in order to merge the three high level customer attributes (the customer attributes and cell information from the mobile network are not hashed).
- Summarizing the hashed IMEI and the high-level customer attributes and cell information on paths and removing all of the IDs.
- Applying threshold values when storing data: ensuring there are a minimum number of 5 hash IDs with the same zip code per radio cell per hour. If the minimum number is not achieved, the zip code length is reduced from the end until the required threshold value is reached. If this is not sufficient, the zip code is completely omitted. Similarly, if zip code information for a movement path does not meet the criteria above in a segment of the path, the zip code information for the whole movement path (in all segments of the path) is removed.

- Applying threshold values when storing data: minimum number of 5 hash IDs per radio cell per hour. If the minimum number is not achieved, no data for this cell and hour is committed to the long-term storage facility.
- Mapping these paths to geographical information with the help of probabilities (geo-mapping to streets and intersections).
- Extrapolating the paths with regional and case-specific algorithms.
- Adhering to threshold values prior to forwarding anonymous data to Motionlogic– minimum number = 30. The case numbers in the result reports are reviewed prior to forwarding to Motionlogic. If a result value does not reach the number 30, the data records are not forwarded.
- Applying a time threshold after a SALT change occurs.

This is shown diagrammatically below



HSM: High Security Module
RGN: Random Number Generator

IMSI: International Module Subscriber Identity
IMEI: International Module Equipment Identity

Alter: Altersgruppe in 10 Jahresschritten
Gen.,: Geschlecht
PLZ: Postleitzahlen

Technical robustness of this solution is maintained by

- Using a secure **cryptographic hash function (SHA2)** for the pseudonymisation of IMSI and IMEI.
- Applying suitably chosen SALT values prior to use of the SHA2 function for IMSI/ IMEI hashing where the SALT.
    - is sufficiently long that it is not practical to determine its value via a brute force with reasonable efforts and within reasonable time and
    - is formed sufficiently randomly and unpredictably (cannot be guessed or reproduced)
- Applying k-anonymisation techniques to ensure that at least 5 distinct hash IDs per radio cell per hour are achieved and at least 5 identical zip codes per radio cell per hour are achieved for data being stored.
- Applying k-anonymisation techniques to resulting analytics shared with the end customer (Motionlogic) to ensure that attribute combinations meet a minimum threshold of 30 occurrences.

3.7 DIFFERENTIAL PRIVACY

Where it is possible to request the result of analytics using different query criteria it may be possible to compare different result sets to determine something about an individual without the direct disclosure of personal information about that individual.

As an example, a hypothetical IoT smart metering service has 10,000 customers, and via a query API allows a third party to request the average energy consumption for each customer's home over an arbitrary set of geographical areas and time periods. The query mechanism has been designed to secure privacy by making sure it will only respond if at least 200 customers have been matched and furthermore only returns the average energy consumption. At first inspection this apparently protects the personal information of any individual customer by use of aggregation and control of the selected data set size. Say though that a malicious third party (however unlikely this is) knows the home location for all 10,000 customers and can formulate a query in such a way as to get the average energy consumption for both the full customer set and the subset of 9,999 customers. The malicious third party can then determine the energy consumption of the final customer and from queries over various time periods can make a reasonable deduction over whether the remaining customer is home, what the customer pays annually for energy, roughly what their annual income might be, etc.

Similar 'differential based attacks' against other services could result in the ability to determine whether a particular user has a particular disease such as heart disease, cancer, HIV or their life expectancy. Other attacks could determine personal data such as user height, weight, age, gender, sexuality, religion, etc.

'Differential Privacy' aims to protect privacy by ensuring an attacker, even in possession of information about every other individual person in a data set, is no more likely to determine a fact about the remaining person in the data set than if they were to flip a coin. The methodology to address the privacy issue was developed by Dr Cynthia Dwork when working for Microsoft Research and is published in the paper 'Differential Privacy'[36]. Differential Privacy suggests various techniques that apply 'noise' statistically to a data set in such a way that the overall results are essentially unaffected, but sampling subsets of data does not disclose personal data about any individual subject.

Differential Privacy has strong support from:

- Apple[37] "... One example is our pioneering use of Differential Privacy, where we scramble your data and combine it with the data of millions of others. So we see general patterns, rather than specifics that could be traced back to you.";
- Orange, described in the blog post "Differential Privacy or how to anonymize datas while managing its usage"[38];
- Google[39] "Often, the training of models requires large, representative datasets, which may be crowdsourced and contain sensitive information. The models should not expose private information in these datasets. Addressing this goal, we develop new algorithmic techniques for learning and a refined analysis of privacy costs within the framework of differential privacy.";
- Microsoft who have developed a framework called 'PINQ'[40] (Privacy Integrated Queries) which uses differential privacy to guarantee the results of a database query will preserve user privacy.

Additional information about differential privacy is available from the Harvard University Privacy Tools Project at https://privacytools.seas.harvard.edu/differential-privacy

[36] The paper was presented at ICALP 2006 'The 33rd international conference on Automata, Languages and Programming' and is also published by Microsoft at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf

[37] See https://www.apple.com/uk/privacy/

[38] See https://recherche.orange.com/en/differential-pivacy-or-how-to-anonymize-datas-while-managing-its-usage/

[39] See https://ai.google/research/pubs/pub45428

[40] See https://www.microsoft.com/en-us/research/wp-content/uploads/2010/09/pinq-CACM.pdf

3.8 TRANSPARENCY, CHOICE AND CONTROL

Widely accepted privacy by design principles stress the importance of transparency, choice and control for storage and processing of personal data. As mentioned above this can be difficult for certain types of IoT devices or service offerings as they may not have a user interface through which the end user can intuitively accept and express consent.

Some of the recommended solutions for gathering consent include:

- For devices which have a suitable user interface it may be practical to gather the user's consent directly either when the device is installed or when it is used;
- For devices which have no user interface but are 'paired' with another device such as a smart phone application it is possible to handle the consent process on that other device;
- For devices which are shared, and subject to the use case e.g. assisted driving solutions for cars, if consent is being sought it may be necessary to obtain the consent from the specific user at that time e.g. if CCTV image or other personal data could be uploaded to a cloud service and even block access to the service if consent is not provided e.g. a connected car not being allowed to start unless consent provided;
- For devices without a suitable user interface, or companion application, or where the user is unable to provide consent interactively it may be necessary to provide some sort of offline opt "in" or "out" service (such as a registration card scheme).

Greater emphasis on consent as part of the European Union's GDPR, has led to the following trends:

- Opt-in by default is being replaced by specific opt-in for many services. This is clearly more difficult in the case there is not a user interface or companion app for the IoT device;
- Service designers are recording what specific opt-ins have been made, and exactly when the consent was provided. Therefore, when processing a data set it may be necessary to check that a relevant opt-in has been obtained within a stated period of time;
- 'General' opt-ins are being replaced by more specific opt-ins, particularly for topics such as third-party marketing partnerships. Therefore, it is a good idea to categorise the opt-ins and match according to the type of analytics or other processing or data sharing that will be taking place.

For further information regarding regulatory requirements on consent see the associated GSMA publication 'Assessing regulatory requirements of privacy management for members offering IoT services using personal data[41]'.

[41] https://www.gsma.com/iot/iot-knowledgebase/assessing-regulatory-requirements-of-privacy-management-for-members-offering-iot-services-using-personal-data/

CASE STUDY : KDDI Privacy Policy Manager

There is an increasing volume of information provided by IoT devices and smartphones in the form of "big data" which can be utilised for various purposes including processing by AI technology.

The Japanese Government is promoting "Ultra Smart Society" initiatives based on science and technology as "Society 5.0". This "Data Driven Society" is an ideal in which both dynamic and static data is created, collected, processed and distributed to generate data analysis to solve issues and support new initiatives. Examples include; open data in the cyber space and the management of agriculture and infrastructure, machine-to-machine and personal data.

In order to realise such a society, it is essential to improve a data distribution platform with a key focus on security and privacy.

PERSONAL DATA PROTECTION

Due to the emergence of smartphones and IoT devices, it is now technically much easier to collect personal and IoT data and to create innovative new services and customised services for each user. In addition, it is expected that personal data is useful as valuable input used to solve wider social issues.

One of the issues when using personal data is that such data can include private information. Appropriate solutions are needed for personal data protection; in particular addressing data leakage and unauthorised use.

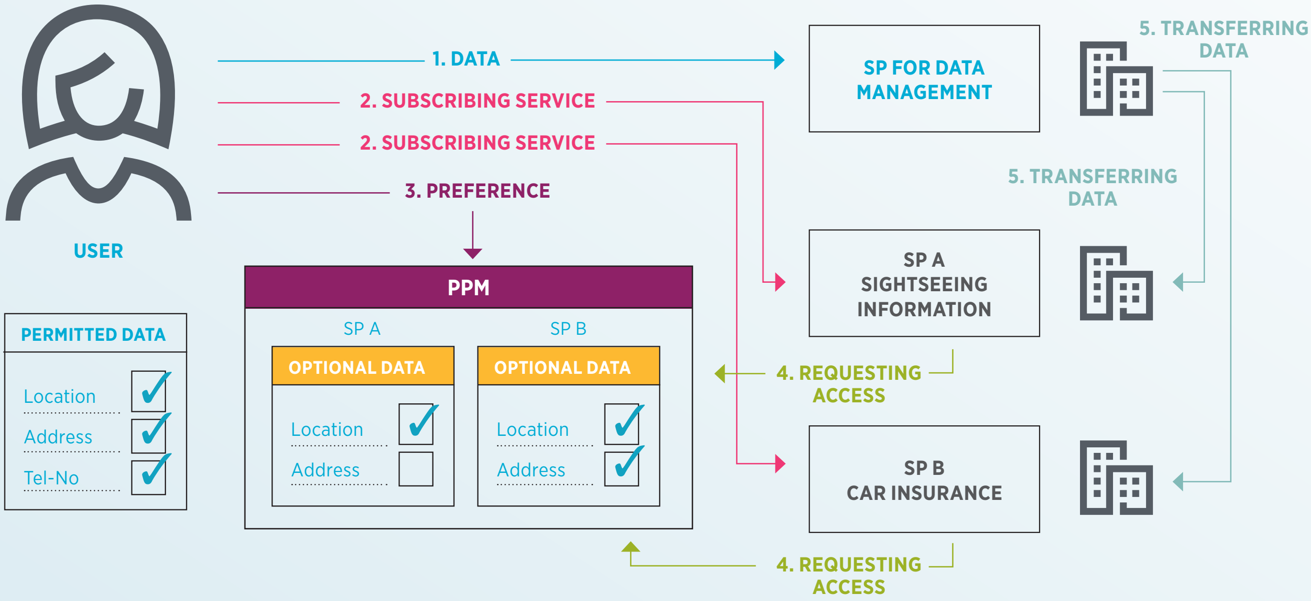
To avoid data leakage, it is necessary to employ mechanisms to block unauthorised access and provide eavesdrop-resistant features (e.g. encryption) on devices collecting information, as well as in the platforms for storing data (e.g. the "cloud").

KDDI's perspective is that it is necessary to handle personal data distribution and utilisation based on the user's intention and it is therefore important to create a mechanism to support this.

TECHNICAL FRAMEWORK TO HANDLE PERSONAL DATA

KDDI Research has developed a framework called "Privacy Policy Manager" (PPM) which enables users to manage appropriate use of their personal data. Current service providers inform users of Terms and Conditions including information to be collected and the purpose of its use. Although in general, within this framework, collection of information is based on users' consent, PPM makes it possible to select items of personal information to be provided and to which service providers the information can be released. Further, PPM visualises the history of data provision.

The figure (below) shows the flow of information when using PPM. Based on its preference table, PPM knows to which service provider the user's data should be distributed after collection.



Data flow and preference in PPM

The details of the flow are as follows;

i. User selects data items (In this case, location, address and phone number are selected) and allows the data management party to provide them. (The data management party collects and manages users' personal data.)

ii. User selects services by service providers (In this case, SP A and SP B in the above figure) which are seeking to use their data. In the process, the user sees details of the data items each service is requesting.

iii. User registers their data distribution preference for each service at PPM.

iv. A service provider accesses the PPM to get permission to access to each user's data. Based on the users' preference, PPM provides tokens for data access to each service provider.

v. Service provider presents the token to the data management party to receive data provided by user.

This framework makes it possible for users to control the data once provided to the data management party and share it to other service providers and other services. With PPM, it is possible for users to manage data provision towards multiple service providers and also to see how data is used by each service provider.

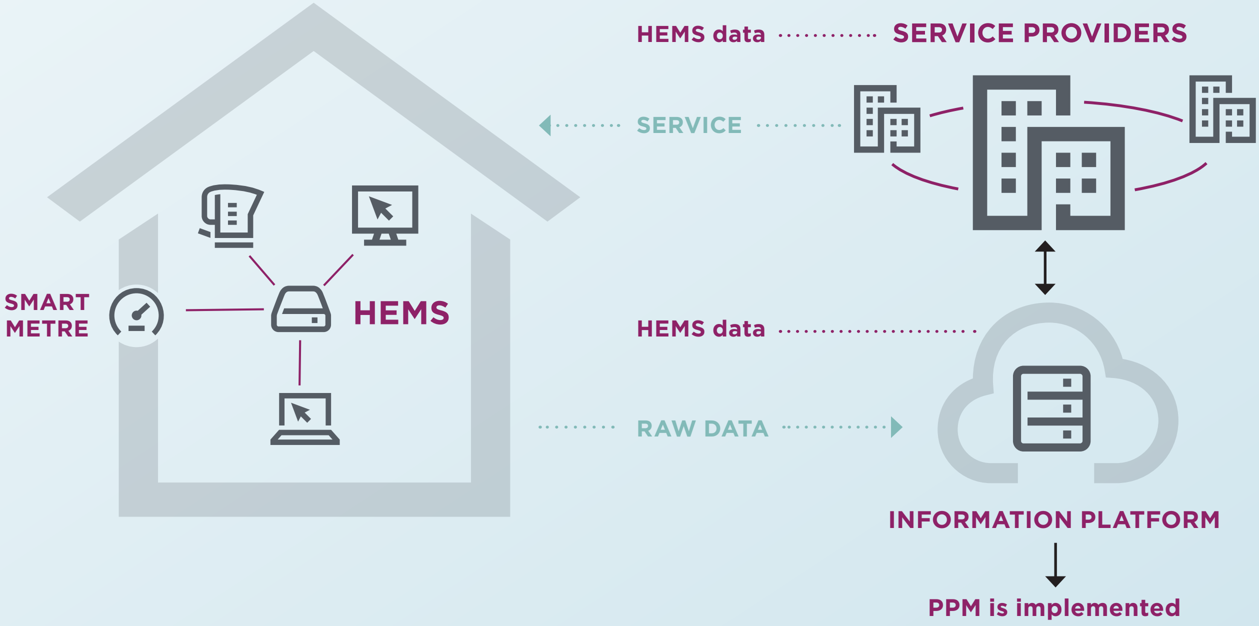
EXAMPLES OF PERSONAL DATA USE BY PRIVACY POLICY MANAGER

KDDI Research has several examples of PPM field deployments which use personal data safely and effectively as below;

HEMS (Home Energy Management System)

The large-scale HEMS information project was held in fiscal year 2014 and 2015. The purpose of the project was to analyse large volumes of electric power consumption with the aim of improving the efficiency of home energy usage. KDDI's involvement in the project was to analyse data from 14,000 households that opted in to the trial.

The HEMS data analysis supported multiple use cases including; a 'daily life assistant', customer offers and customer advice on energy saving. These use cases were delivered by multiple service providers processing the energy consumption data coming out from each household. In the trial, information management parties including KDDI received HEMS data from each household and passed them to service providers using HEMS data for their service provision. (Figure below)



Home Energy Management System

KDDI's PPM provided participating households with functions to select data items to be provided to service providers and check data usage history.

Electric receipt

Another project, "electric receipt", was held in 2016 selected by the Japanese Ministry of Economy, Trade and Industry for IoT business to create new industry models based on consumer purchase data. The purpose of this trial was to improve standardisation related to consumer purchase history and show how this purchase history can be used for the consumer benefit. When a consumer buys goods in a shop, the purchase data is stored as an electric receipt in a cloud-based data store. The consumer can download their own purchase data to a smartphone app.

KDDI identified several cases using electric receipt through PPM; recommendation of cooking recipes, questionnaires and answers about goods purchased and loyalty points. PPM was provided in the smartphone app for consumers to check electric receipts and data distribution. Based on the preferences set by the consumer, the app enables access to purchase history and attributes (age, address, etc.) to data distribution parties. The app also allowed the granularity of data to be adjusted, for example changing from release of specific product names to types like vegetables, sweets, etc.

TOWARDS PERSONAL DATA UTILISATION

KDDI sees that there is a need to establish a common framework of personal data provision for various users who have different perspectives on sharing personal data. This will be the case even considered within the need to align with each country's regulations and rules related to privacy protection. KDDI's work to deliver PPM has confirmed that users are keen to enjoy the benefit of a common framework for managing the relationship between use of highly convenient services and managed sharing of personal data.

3.9 DATA ERASURE

Aside from the familiar case of users "opting in" to services as described above, users may be entitled in certain jurisdictions to have their account 'forgotten' and personal data 'erased'.

In supporting these requirements, and particularly where personal data forms part of a data set used in big data analytics and machine learning, the following best practice approaches can be implemented:

- If removing records or data would impact obligations such as maintaining official records for tax purposes, public interest, or legal obligations the records or data could be moved to an archival system. Often there will be a period of time that records need to be retained for compliance purposes after which they can and should be deleted or the personal data erased. Systems should be coded in such a way as to handle different compliance and retention requirements depending on country or region;
- Personal data which is not required for government or regulatory compliance can be de-personalised using techniques such as anonymisation. It may be feasible to simply delete records from a database or if this affects other processing the personal data content can be erased by overwriting with null or randomised (i.e. anonymised) values;
- It is important to have traceability between any items of personal data and the user account which that relates to so that all the data pertaining to the user can be identified and deleted or erased if appropriate;
- It is important to record information such as account creation date, last login date, contract start & expiry date so that erasure & retention policies can be enforced by the system;
- If personal data has been incorporated into data sets, they may also need to be updated through erasure of whole records or the fields containing the personal data, there may be a need to regenerate analytics if there is any of the personal data included in the output;
- If personal data was incorporated in data sets for training machine learning it is possible the machine learning models would react to, memorise or even potentially output personal data and therefore might need to be retrained.

3.10 DATA ORIGINATION TRACEABILITY

The previous section on data erasure and the right to be forgotten handles individual cases where customer details are selectively removed from a direct service. There are broader issues however particularly where data sets are sourced from third parties as resulting analytics/ machine learning would then be based on data which is derived from those external data sources.

Third party data, whether open source or commercial:

- May be subject to specific licensing conditions from the publisher which limit certain applications, distribution rights, use or retention periods, etc.;
- A whole data set might be withdrawn by a supplier for various conditions, in which case any downstream analytics, machine learning, result publishing or onward data distribution needs to be reviewed or potentially re-engineered to work without that data set;
- If an external data set also includes data about an individual (even if they have previously consented for their data to be included in the data set) the permission could be revoked and this would need to be dealt with in any system receiving that data.

It is important therefore to maintain traceability over the origins of data as datasets are joined together, analysed either using regular algorithms or machine learning, presented as results or even distributed to third parties.

3.11 ACCOUNT SECURITY

User account security is increasingly subject to hacking attempts. Various methods can be considered to improve security beyond traditional username and password schemes. One example is multi-factor authentication which introduces an extra layer of account security, replacing or in addition to usual username/email plus password combinations.

If data belonging to a user is being stored on a big data platform, or analytics can be obtained from that platform, there is a benefit to employing multi factor authentication to prevent third parties compromising the privacy and security of one or more user accounts. This is especially important as the fines for data breaches possible with new privacy rules are now potentially much higher. In the case of EU GDPR fines can reach up to 4% of the company world-wide annual turnover of the preceding year[42].

[42] Art 83(5) EU GDPR – General conditions for imposing administrative fines

Typically, the extra security is in the form of a physical device such as a hardware token, or a dedicated software application that the user runs on a physical device. The Mobile Connect service[43] provides a solution for multi-factor authentication based on the mobile phone account of the user, secured using the mobile network and SIM card.

Mobile Connect Authenticate requires the operator, on receiving a Mobile Connect API request from an application, to confirm that the end user is in possession of their mobile device (LoA2) and/or has entered a secret that they know (PIN) when prompted on their mobile device (LoA3).

Mobile Connect Authentication in four steps

1

The service provider indentifies the serving operator to authenticate an end user

2

The operator challenges the end user by asking to prove possession of the mobile phone and optionally enter a pin code

3

The end user confirms the possession of one or both creditional factors on the mobile phone

4

The operator provides authentication success or failure response back to the service provider

The authenticator used will depend upon the individual operator and the **Level of Assurance** requested by the application. The table below illustrates the different authenticators and the Level of Assurance (LoA) they offer.

AUTHENTICATOR	DESCRIPTION	LOA2	LOA3	LOA4
Seamless Authentication	Authentication is automatically handled by the operator is the user is connected via the operator network.	✓		
SMS+URL	The end-user verifies themselves by clicking on a link in an SMS.	✓	✓	
USSD	A USSD session is initiated allowing the end-user to verify themselves.	✓	✓	
SIM Application Toolkit	A SIM Toolkit session is initiated allowing the end-user to verify themselves.	✓	✓	✓
Smartphone application	A native application that allows the end-user to manage their verification.	✓	✓	✓

[43] https://www.gsma.com/identity/mobile-connect

Mobile Connect has been designed with privacy focus at the core, preserves citizens' trust and is aligned with government and regulators priorities. For example, in keeping with worldwide privacy requirements available worldwide (e.g. EU General Data Protection Regulation), Mobile Connect supports the principle of privacy by design, seeking to ensure the services and an individual's identity attributes are used in secure, privacy respective and protective ways.

Pseudonymous Customer Reference

The Mobile Connect Pseudonymous Customer Reference (PCR for short) reduces the risk of "link-ability" to the data subjects.

A PCR is used to ensure that the end-user's privacy is protected while the developer can be confident that a PCR represents an actual end-user. Developers can then provide access to online services using the PCR and may request additional information about a user with their consent. The PCR is a persistent but non-personal identifier unique to the mobile phone number and application, other Mobile Connect-enabled applications and services will not be able to use a PCR which doesn't belong to them for user authentication. The PCR also mitigates the risk of data linkage because multiple applications receive different values for the PCR for the same mobile phone number.

3.12 CROSS BORDER DATA FLOWS

Some jurisdictions apply restrictions to the flow of personal data "cross border" and may require it to be located in a specific country or region. The GSMA has recently published the report 'Cross-border data flows, realising the benefits of removing barriers'[44] which provides full details of the type of restrictions and explains the benefits to citizens, society and organizations for data to freely flow across borders. For IoT services, with a global customer base or devices which are not confined to remaining in a single country or region, these restrictions may be addressed with a number of technical solutions. Note that GSMA does not endorse moves to restrict cross border data flows or recommend data is located in specific countries or regions except as required by regulation.

Some of the relevant approaches include:

- Deploying servers in specific countries or regions as required to contain user data in the required geographical area. This generally leads to increased costs, potentially lower service speeds, and negative downstream effects on consumers and businesses;

[44] https://www.gsma.com/publicpolicy/wp-content/uploads/2018/09/GSMA-Cross-Border-Data-Flows-Realising-benefits-and-removing-barriers_Sept-2018.pdf

- Having the ability to route users, data and services to the servers that hold the user data. This can potentially use solutions such as Amazon's Route 53 service[45] which allows geographical routing based on the IP address of network requests over the Internet[46]. However even if using a 'geo IP' service such as this it is still usually necessary to have the capability of routing customer data to the correct country or region to cover cases where the IP routing has not worked with 100% effectiveness or where the user or device are outside the required country or region;
- 'Load balancing' and 'failing over' to servers also located in the required country or region. Again, this can lead to increased cost and complexity of service delivery.

Data location restrictions can have an impact on the analytics and services delivered for a global IoT customer base. There is often an advantage in processing, analysing and the use of machine learning on a data set comprising data collected across the globe. For example, in the area of health it is possible to determine conditions which affect specific ethnicities in particular ways; similarly, automotive manufacturers typically sell the same product globally; and many supply chains involve transportation internationally across different regulated borders.

Data transfers outside of the country or region can reasonably still take place if contractual and/ or technical measures are put in place.[47] The following additional approaches, covered in detail elsewhere in this guide, are useful to assist with cross border restrictions:

- Use of anonymisation and pseudonymisation techniques to make data less personally identifiable;
- Aggregating data so that it is now about a group of users or devices rather than an identifiable individual;
- Using strong encryption when transferring data between countries/ regions and when storing data (at rest);
- Gaining user's informed consent for data storage & processing;
- Using techniques such as 'k-anonymisation' and 'differential privacy' to minimise privacy risks.

[45] Further information on Route 53 is available at https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html

[46] Note that IP address based geographical routing is not 100% accurate because IP addresses are not strictly allocated along geographical lines. Databases are typically around 98% accurate, but routing can also be incorrect for traffic using corporate WANs, VPNs, etc. Even so this can be a help to send the large majority of users to the 'correct' system.

[47] See also GSMA -Cross border data flows – realising the benefits and removing barriers https://www.gsma.com/publicpolicy/wp-content/uploads/2018/09/GSMA-Cross-Border-Data-Flows-Realising-benefits-and-removing-barriers_Sept-2018.pdf

3.13 HOSTING

It is important to establish an appropriate hosting environment to protect the privacy of users. This is the case whether the hosting uses ‘on premises’ equipment, traditional hosting services e.g. dedicated servers, or ‘cloud’ based hosting. Where big data and machine learning are concerned a reasonable expectation is there will be large amounts of data distributed across multiple processing or storage nodes.

Hosting can also be affected by cross border data flow requirements, as presented in the previous section. Some jurisdictions may restrict data to flow freely from one country to another and as such require to be hosted within the borders. For a more thorough description on existing restrictions and why it is best for citizens, organisations and government to let data flow across the border refer to GSMA paper ‘Cross-border data flows, realising the benefits of removing barriers’.

Key best practices in selecting and implementing hosting therefore include:

- The support for ‘strong’ data encryption at rest at the various nodes and in transit between nodes;
- Understanding issues around data localisation/ cross border data flows and the requirements for storing data;
- Secure cleansing and/or destruction of decommissioned storage devices e.g. hard disk and solid-state drives so that discarded equipment does not risk disclosure of personal data. This is even in the case that data is encrypted at rest because having data secured now according to best practice does not mean it will necessarily be secure in the future;
- First line defences e.g. network firewalls and web application firewalls[48] protecting servers from external attacks that could lead to personal data loss;
- Secure administration functionality ensuring the servers and storage are well protected against hackers;
- Secure backup facilities so that there is no risk of ‘back-doors’ being opened up through normal operational procedures or loss of service if a system is compromised;
- Ongoing auditing of servers for operating system and application software patches and reporting the need for updates to administrative staff.

[48] See https://www.owasp.org/index.php/Web_Application_Firewall

3.14 OUTSOURCING

There are some key considerations regarding outsourcing of analytics, machine learning and service development and/or systems operations services and the protection of privacy. Many privacy frameworks use the principle of 'accountability'[49]. This is effectively a requirement on the data controller and processor to protect data privacy when it is passed on to third parties. The obligation flows down to data subjects and in principle across borders. However, implementation methods and best practice differs. Aside from contractual protections the following practices may be considered:

- During the development of analytics and machine learning it will be necessary for data scientists to explore and analyse data sets in various ways. Often, during initial stages of development, it will be necessary to create very broad data sets to identify the subset of attributes that are actually useful to a given problem. The various techniques of anonymisation, pseudonymisation and aggregation can be employed to protect personal information in the input and output data sets used during development, especially if some of this work is outsourced outside of the home country/ region;
- Any data set used for development, including the input data set for machine learning training, should be securely encrypted when sharing with outsourcing partners with tightly controlled access to encryption & decryption keys and use of strong credentials and access control;
- Outsourcing partners should delete any copies of data sets once their project need for the data is complete unless there are justifiable purposes for which it must be retained;
- A trained machine learning model might 'learn' to recognise, act on or output certain personal data under circumstances that it may be impossible to understand and so it is important to limit access to such a trained model;
- System access controls should be allocated appropriately to the need of the outsourcing partners so that access to systems and data is restricted to just those people who need it;
- Other protections such as listed above under hosting should apply to any systems provided by an outsourced service company;
- Outsourcing partners should also implement all relevant protections against OWASP Top 10 risks[50] for any parts of an analytics, machine learning or service they are responsible for.

[49] See https://www.gsma.com/publicpolicy/wp-content/uploads/2018/09/GSMA-Regional-Privacy-Frameworks-and-Cross-Border-Data-Flows_Full-Report_Sept-2018.pdf
[50] See https://www.owasp.org/index.php/OWASP_Top_Ten_Cheat_Sheet

04 Related resources

Privacy and the related topic of security are substantial topics, and this guide should be consulted along with the following complementary resources:

01. GSMA publication “Assessing regulatory requirements of privacy management for members offering IoT services using personal data” https://www.gsma.com/iot/iot-knowledgebase/assessing-regulatory-requirements-of-privacy-management-for-members-offering-iot-services-using-personal-data/

02. GSMA IoT Security Guidelines and Assessment.
https://www.gsma.com/iot/future-iot-netwoks/iot-security-guidelines/

03. GSMA Mobile Privacy Principles.
https://www.gsma.com/publicpolicy/mobile-privacy-principles

04. GSMA Mobile Privacy and Big Data Analytics.
https://www.gsma.com/publicpolicy/mobile-privacy-big-data-analytics

05. The Open Web Application Security Project which covers best practices across many areas related to privacy and security. https://www.owasp.org/index.php/Main_Page

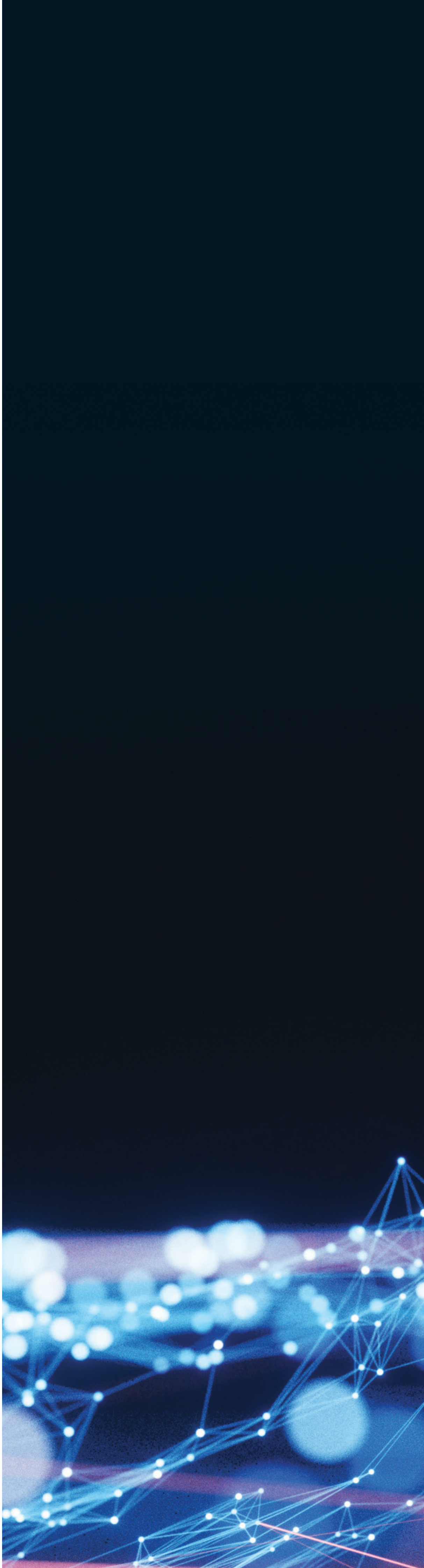
a. Top 10 Privacy Risks.
https://www.owasp.org/index.php/OWASP_Top_10_Privacy_Risks_Project

b. Top 10 Security Vulnerabilities.
https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project

05 Definitions

TERM	DESCRIPTION
Encryption	A method of securing data so that it can only be accessed by a person, system or organisation that is in possession of a key which can be used to decrypt the encrypted data. Often public/private key encryption is used (e.g. RSA) so that anyone in possession of a public key can encrypt data but only the possessor of the private key can decrypt the data. See also https://searchsecurity.techtarget.com/definition/encryption
Hashing	A method of converting a data item to a representation of the bits of data such that it is difficult to reconstruct the original data item. In this document hashing is proposed for the encoding of personal data using an SHA2 based algorithm such as SHA-256 or higher. See also https://techterms.com/definition/hash
Initialisation Vector	In encryption an Initialisation Vector is important to protect the encryption process from attacks against multiple encryption sessions. This adds randomness to the start of the encryption process so that it is impossible to learn anything from the encryption of the same data or subsets of the same data from subsequent encryption sessions. See also https://whatis.techtarget.com/definition/initialization-vector-IV
Salting	'Salting' is a method of adding random data to original data so that it is more difficult to decrypt or 'un-hash' by an attacker looking for common data. Salting is recommended when MSISDNs, IMSIs or related identifiers are hashed for storage by analytics platforms. The process should also be used for encryption of any other sensitive data. See also https://auth0.com/blog/adding-salt-to-hashing-a-better-way-to-store-passwords/
VPN	Virtual Private Network – a way to extend private networks to external devices using a 'tunnel' which typically operates across the Internet using end to end encryption. See also https://www.webopedia.com/TERM/V/VPN.html

For more information please visit:
www.gsma.com/IoT



GSMA HEAD OFFICE
Floor 2
The Walbrook Building
25 Walbrook
London EC4N 8AF
United Kingdom
Tel: +44 (0)20 7356 0600
Fax: +44 (0)20 7356 0601